A fully vector quantised self-excited vocoder

Redwan A. Salami, Lajos Hanzo and Derek G. Appleby

Department of Electronics and Computer Science
University of Southampton
Southampton SO 9 5 NH
UK

ABSTRACT

A computationally efficient stochastic vector quantisation (VQ) scheme for the quantisation of the LPC parameters with less than 25 bits is proposed. The complexity of the self-excited vocoder (SEV) is reduced by employing fast algorithms based on an efficient representation of the innovation sequence. When deploying the suggested VQ scheme in the improved SEV, good communications speech quality has been resulted at 4.8 Kb/s. When the bit rate is increased to 6.8 Kb/s, near-toll quality speech has been achieved.

1. Introduction

Low bit rate speech coders such as the Self-Excited Vocoder (SEV) [1-2] and Code-Excited Linear Prediction (CELP) coders [3] went through a quick evolution in recent years. Researchers are striving to simplify these high-complexity speech coding algorithms to achieve their real-time implementation. At bit rates around 4.8 Kb/s, the quantisation of LPC parameters with less than 25 bits becomes crucial to allow more bits to be used for quantising the excitation signal. We suggest in this paper an efficient and practical stochastic vector quantisation scheme for the quantisation of the LPC parameters, using less than 23 bits while maintaining low spectral distortion. This vector quantisation (VQ) scheme is deployed in a proposed fast SEV scheme for speech coding at 4.8 Kb/s. The computational efficiency of the proposed SEV is due to the simplified innovation sequence representation, as it will be shown in Section 2. The description of the stochastic VQ scheme is given in Section 3.

2 Self-excitation Algorithms

2.1 Basic SEV

Let us consider the basic SEV scheme depicted in Fig.1, where A(z) is the all-pole LPC analysis filter and $W(z)=1/A(z/\gamma)$ represents the ubiquitous error weighting filter (the LPC synthesis filter $1/A(z)$ cascaded with the filter $A(z)/A(z/\gamma)$). The transfer function of the long term prediction (LTP) synthesis filter with delay γ_1 and gain β_1 is given by $P(z)=1/\left(1-\beta_1 z^{-\gamma_1}\right)$. The cascaded filter combination P(z), W(z) is excited by the innovation sequence $v_2(n)$, which is selected from the buffered past history of the excitation to produce the perceptually best synthetic speech. Since the LTP residual can be considered to be a random stochastic process, we judiciously initialise the $v_2(n)$ buffer from a zero mean, unit variance Gaussian random process at the beginning of a coding session. Then we continuously find in the $v_2(n)$ buffer that particular delay γ_2 with the associated gain β_2, which minimises the perceptually weighted error between the original speech frame and the synthesized one.
The pitch predictor parameters (β_1,γ_1) can be computed directly from the short term prediction (STP) residual r(n). However, a large improvement can be scored, if the pair (β_1,γ_1) is computed inside the optimization loop. Let $h(n)$ denote the impulse response of the weighting filter $W(z)=1/A(z/\gamma)$. Then the weighted error between the original and synthesised speech is given by:

$$e_w(n)=s_w(n)-\hat{s}_w(n) \tag{1}$$

and

$$\hat{s}_w(n)=v_1(n)*h(n)+\hat{m}(n), \tag{2}$$

where the convolution * is a memoryless process and $\hat{m}(n)$ is the memory of the filter W(z) due to its initial state from the previous speech frame's contribution. Therefore:

$$e_w(n)=x(n)-v_1(n)*h(n), \tag{3}$$

where $x(n)=s_w(n)-\hat{m}(n)$ is the weighted speech after subtracting the memory contribution of the filter W(z). The input $v_1(n)$ of the filter W(z) is given by:

$$v_1(n)=\beta_2 v_2(n-\gamma_2)+\beta_1 v_1(n-\gamma_1), \tag{4}$$

therefore:

$$e_w(n)=x(n)-\beta_1 v_1(n-\gamma_1)*h(n)-\beta_2 v_2(n-\gamma_2)*h(n). \tag{5}$$

By defining the signals $y_1(i,n)$ and $y_2(i,n)$ as below:

$$y_1(i,n)=v_1(n-i)*h(n)$$

$$y_2(i,n)=v_2(n-i)*h(n), \tag{6}$$

the weighted error becomes:

$$e_w(n)=x(n)-\beta_1 y(\gamma_1,n)-\beta_2 y(\gamma_2,n). \tag{7}$$

Jointly optimising the pairs $(\beta_1\gamma_1)\wedge(\beta_2\gamma_2)$ to minimise the mean squared error (MSE) leads to:

$$E_w=\sum_{n=0}^{N-1} x^2(n)-\beta_1\sum_{n=0}^{N-1} x(n)y_1(\gamma_1,n)$$

$$-\beta_2\sum_{n=0}^{N-1} x(n)y_2(\gamma_2,n). \tag{8}$$

According to this optimum approach , the error E_w should be computed for all possible combinations of γ_1 and γ_2, which results into excessive computational demand. Therefore a suboptimum two stage approach is pursued, in which we first set $v_2(n)=0$, compute (β_1,γ_1) and then find (β_2,γ_2) in a second stage. Whence the weighted error becomes:

$$e_w(n)=x(n)-\beta_1 y(\gamma_1,n). \tag{9}$$

By setting $\partial E_w/\partial\beta_1=0$, we get:

$$\beta_1=\frac{\sum_{n=0}^{N-1} x(n)y_1(\gamma_1,n)}{\sum_{n=0}^{N-1}[y_1(\gamma_1,n)]^2}, \tag{10}$$

$$E_w=\sum_{n=0}^{N-1} x^2(n)-T(\gamma_1) \tag{11}$$

where

$$T(\gamma_1)=\frac{\left(\sum_{n=0}^{N-1} x(n)y_1(\gamma_1,n)\right)^2}{\sum_{n=0}^{N-1} y_1^2(\gamma_1,n)}. \tag{12}$$

To minimise the MSE the term $T(\gamma_1)$ in Eq (12) has to be computed for the range of $N<\gamma_1<(L-1)$, and that particular delay is chosen, which maximises $T(\gamma_1)$, ie minimises E_w in Eq (11).

CH2673-2/89/0000-0124 $1.00

The LTP delay γ_1 is restricted to $\gamma_1 > (N-1)$, where N is the excitation frame length, so that the impulse response of the combined filter P(z)W(z) is the same as that of W(z) for n<N. Therefore P(z) is not considered while optimising the excitation parameters β_2 and γ_2.

The number of multiplications needed to compute the convolution $y_1(i,n)$, n=0...(N-1) for a given i is N(N+1)/2. A dramatically cut computational load is resulted by employing the following relation for the computation of $y_1(i,n)$ for the range $N < i < (L-1)$

$$y_1(i,0) = v_1(-i)h(0), \tag{13}$$

$$y_1(i,n) = y_1(i,0) + y_1(i-1,n-1), \quad n = 1\ldots(N-1).$$

The procedure used to determine the LTP parameters β_1 and γ_1 can now be repeated for the computation of the pair (β_2,γ_2), with the following modifications:

$$v_1(n) = v_2(n) + \beta_1 v_1(n-\gamma_1), \tag{14}$$

$$e_w(n) = x_2(n) - v_2(n)*h(n), \tag{15}$$

$$x_2(n) = x(n) - \beta_1 y_1(\gamma_1,n), \tag{16}$$

where now $x_2(n)$ is the weighted speech after subtracting the contributions of the memories of both P(z) and W(z).

The LTP parameters (β_1,γ_1) and the innovation sequence parameters (β_2,γ_2), along with the LPC filter parameters have to be quantised and sent to the decoder, where they are used to reproduce the synthetic speech $\hat{s}(n)$.

2.2 Efficient SEV Algorithms

Although the complexity of the original SEV is already realistic for real-time implementations, we further increase its computational efficiency, without seriously effecting the perceived speech quality.

From our experience with the CELP codec, we have found that the codebook entries can be well represented by a reduced number of nonzero pulses, whose positions can be regularly spaced. Let us hence represent the innovation sequence $v_2(n)$ by the help of a sequence, having non-zero values only at positions n=D.i, i=0...(N/d)-1, where D is a decimation factor in the range $2 \le D \le 5$. The rest of the $v_2(n)$ samples are set to zero. Accordingly, the delay γ_2 is now restricted to be the multiple of D, therefore the range of γ_2 values is reduced by a factor D.

Although $v_2(n)$ contains now zero samples, its weighted counterpart $y_2(n)=v_2(n)*h(n)$ does not necessarily contain zeros. To preserve zero samples in the computation of the weighted error and therefore keep the computational load low, we use the autocorrelation approach in the MSE formulation. Accordingly, the MSE in Eq (11) is now given as

$$E = \sum_{n=0}^{N-1} x^2(n) - \frac{\sum_{n=0}^{N-1} \Psi(n)v_2(n-\gamma_2)}{\Phi(0)\mu(\gamma_2,0) + 2\sum_{n=0}^{N-1}\Phi(n)\mu(\gamma_2,n)}, \tag{17}$$

where

$$\Psi(n) = \sum_{i=n}^{N-1} x_2(i)h(i-n) = x_2(n)*h(-n) \tag{18}$$

and $\Phi(n)$ is the autocorrelation of the weighting filter's impulse response h(n) and $\mu(\gamma_2,n)$ is the autocorrelation of the innovation sequence at delay γ_2. Since $v_2(n-\gamma_2)$ is non-zero only, when $(n-\gamma_2)$ is a multiple of D, the second term of Eq (17) is reduced to:

$$T(\gamma_2) = \frac{\sum_{n=0}^{N/D-1} \Psi(nD)v_d(n-\gamma_2)}{\Phi(0)\mu(\gamma_2,0) + 2\sum_{n=0}^{N/D-1}\Phi(nD)\mu(\gamma_2,n)}, \tag{19}$$

where $v_d(n)$ contains already the non-zero values of $v_2(n)$ only and $\Psi(nD)$ is also confined to the non-zero values only. Therefore, both the computational demand and the buffer length are reduced by a factor D. We refer to this approach in the discussion section as Algorithm 1.

The computation of the autocorrelation $\mu(j,i)$ does not constitute a serious problem, as it is simply updated from delay to delay by taking into the effect of the new first value of $v_2(n)$ and that of the discarded last value, as the computation window is slid along the $v_2(n)$ buffer, i.e.

$$\mu(j,i) = \mu(j-1,i) + v_2(-j)v_2(-j+i) - v_2(-j+N)v_2(-j+N-i). \tag{20}$$

We now proceed to further simplify Eq (19). Recall that

$$x_2(n) = s_w(n) - \hat{m}(n) - \beta_1 v_1(n-\gamma_1)*h(n) \tag{21}$$

where $s_w(n)$ is the weighted original speech expressed as

$$s_w(n) = r(n)*h(n) + m(n). \tag{22}$$

Here m(n) is the output of the error weighting filter in the upper branch of Fig.1, due to its initial state.

Since we are modelling the original speech by the synthetic speech, we judiciously suppose that the memory contributions of the filters in both branches are equal. Therefore from Eq (21) we have:

$$x_2(n) = r(n)*h(n) - \beta_1 v_1(n-\gamma_1)*h(n). \tag{23}$$

Then by substituting Eq (23) into Eq (18) we get:

$$\Psi(n) = r(n)*\Phi(n) - \beta_1 v_1(n-\gamma_1)*\Phi(n)$$
$$= d(n)*\Phi(n), \tag{24}$$

where $\Phi(n) = h(n)*h(-n)$ is the autocorrelation of the impulse response h(n) used in Eq (17), and

$$d(n) = r(n) - \beta_1 v_1(n-\gamma_1) \tag{25}$$

is actually the LTP residual. Therefore, from the last form of Eq (24) a new SEV structure can be soon contrived. By using Eq (24), Eq (19) can be rewritten as:

$$T(\gamma_2) = \frac{\sum_{n=0}^{N/D-1} [d(nD)*\Phi(nD)]v_d(n-\gamma_2)}{\Phi(0)\mu(\gamma_2,0)}, \tag{26}$$

where the second term of the denominator in (19) is neglected by the assumption that $\Phi(nD) \ll \Phi(0)$. This is justified by the fact that h(n) is a sharply decreasing impulse response, therefore the rate of decay of its autocorrelation is even faster. If we define a smoother w(n) as

$$w(n) = \frac{\Phi(n)}{\Phi(0)}, \tag{27}$$

Eq (26) can be reformated as

$$T(\gamma_2) = \frac{\sum_{n=0}^{N/D-1} [d(nD+k)*w(n)]v_d(n-\gamma_2)}{\mu(\gamma_2,0)}, \tag{28}$$

where k=0...(D-1) represents the possible initial grid positions of the decimated sequence d(nD). Now we can derive our new, computationally highly efficient, good perceptual quality RPE-SEV structure, depicted in Fig.2.

The STP residual r(n) is found by filtering a frame of N original speech samples through the inverse LPC filter A(z). The LTP parameters (β_1,γ_1) are computed by minimising the error between the STP residual r(n) and its estimated value $\beta_1 v_1(n-\gamma_1)$. The residual d(n) after removing long term periodicity is filtered by the smoother $w(n) = \Phi(n)/\Phi(0)$, which is the normalised autocorrelation of the error weighting filter's impulse response h(n). The smoothed LTP residual is denoted by $d_s(n)$, which is split into D number of sequences:

$$d_d^{(k)}(n) = d_s(nD+k), \qquad n = 0\ldots(N/d-1). \tag{29}$$

We use these sequences to represent the weighted LTP residual for our excitation matching algorithm. The optimum decimated innovation sequence $v_d(n)$ is given by that (β_2,γ_2) pair, which maximises the term $T(\gamma_2)$ in Eq (28) for all possible combinations of k and γ_2. However, we suggest a sub-optimum approach to keep the computational demand low, in which we first choose that particular $d_d^{(k)}(n)$ sequence,

which contains the highest energy to represent $d_s(n)$ for finding (β_k, γ_k). In our discussion section we refer to this method as Algorithm 2.
Observe that at this stage the smoother w(n) is updated for each new LPC analysis frame. To further simplify the SEV algorithm we can choose w(n) as the normalised autocorrelation of a filter with a low number of fixed prediction coefficients, computed from the long term correlations of speech. The smoother w(n) is sharply decaying, and indeed it can be truncated at $|n| \leq Q$, where $Q \ll N$. If this truncated fixed smoother $w_t(n)$ is shifted by Q samples to the range $0 \leq n \leq 2Q+1$ for the sake of causality, frequency domain analysis shows that it has the transfer characteristic of a low-pass filter. In general, for a decimation factor D, an FIR low-pass filter with a cut-off frequency $f_s/2D$ and length 2Q+1 can be deployed, where f_s is the sampling frequency. The smoother w(n) is obtained by shifting the FIR filter's impulse response to the left by Q samples. This method with a fixed smoother is refered to as Algorithm 3.

3 Vector Quantisation of LPC Parameters

For speech coding with bit rates around 9.6 Kb/s, the log area ratios (LAR) or the line spectrum frequencies (LSF) are usually quantised with 30-40 bits per 20 ms LPC update frame. Below 5 Kb/s encoding rates either the LPC update frame has to be extended (e.g. to 30 ms), or vector quantisation of the LPC parameters with at most 25 bits per 20 ms speech frame has to be deployed.
Conventional vector quantisers [4] use trained codebooks, which usually lack robustness over speakers outside the training sequence. Shoham [5] attempted to exploit the similarities among successive spectral envelopes by employing vector predictive coding , where trained codebooks are needed for the predictor and residual vectors. According to the scheme proposed in [6], an LSF vector is quantised from a codebook containing the previously quantised vectors and then the residual error from this first stage is quantised using a second Gaussian codebook. This approach can be easily implemented but it has the problem of propagating channel errors. In [7], a switched-adaptive method is introduced, which exploits the correlation between adjacent LSF vectors.
In the next section, we propose a practical stochastic VQ method based on an approach published by Atal [8]. In the original approach, the covariance matrix of the LARs is computed from a buffer containing the previously quantised LAR vectors. Then it is decomposed into its eigen vectors and eigen values. This is done for every new LPC frame, which is a computationally rather demanding task. Furthermore, the eigen value solution needs at some stage an iterative algorithm (e.g. the QR algorithm) which makes the processing time data dependent.

3.1 Stochastic VQ of LPC Parameters

In our approach, an LPC parameter vector is quantised using a Gaussian codebook by transforming the uncorrelated codebook entries into vectors having correlations similar to those of the LPC parameter vectors. A vector x of dimension N having jointly correlated components is transformed into a vector u with uncorrelated components by using an orthogonal rotation with an NxN matrix A

$$u = Ax. \tag{30}$$

For a source x whose components are jointly Gaussian, it was shown that the optimal rotation A is given by a matrix, whose rows are the normalised eigen vectors of Γ_x, the covariance matrix of x . This transformation is usually refered to as the Kahrunen-Loeve Transform (KLT), and it can be applied to some extent to non-Gaussian sources [4]. The covariance matrix Γ_x is given by

$$\Gamma_x = E[(x-\bar{x})(x-\bar{x})^T], \tag{31}$$

where E denotes the expectation and $\bar{x} = E(x)$. Γ_x can be decomposed into

$$\Gamma_x = S\lambda S^T, \tag{32}$$

where S is a matrix whose columns are the normalised eigen vectors of Γ_x and λ is a diagonal matrix whose elements are the eigen values of Γ_x. Therefore, the rotated vector u in Eq (30) is given by

$$u = S^T x. \tag{33}$$

It can be shown that the covariance matrix of u is the diagonal matrix λ, which means that u has uncorrelated components. The variances of the components of u in Eq (34) are the eigen values of Γ_x, and their means are given by

$$\bar{u} = S^T \bar{x}. \tag{34}$$

In order to turn the transformed vector u into one with unity covariance matrix and zero means, the following transformation is used

$$u = \lambda^{-1/2} S^T (x - \bar{x}). \tag{35}$$

Then Eq (35) suggests the stochastic vector quantisation method. A vector x is quantised using vectors chosen from a codebook, which contains zero mean, unity variance Gaussian entries through the following transformation

$$\hat{x} = \bar{x} + \beta S \lambda^{1/2} u^{(k)}. \tag{36}$$

Eq (36) is derived directly from Eq (35) with the scalar β introduced to allow more flexibility in matching the powers of x and $\hat{x}$. The mean squared error between the original and quantised vectors x and $\hat{x}$ is given by

$$E_x = (x-\hat{x})^T (x-\hat{x}) = \|x-\hat{x}\|^2 = \|y - \beta\lambda^{1/2} u^{(k)}\|^2 \tag{37}$$

where $\|\ \|$ denotes the Euclidean norm and

$$y = S^T (x - \bar{x}). \tag{38}$$

The optimum codebook gain β is computed by setting $\partial E_x / \partial \beta = 0$.
The codebook of Gaussian vectors $u^{(k)}$ is exhaustively searched for the index k, which minimises the error in Eq (37), and the quantised vector is then computed from Eq (36). The long term covariance matrix Γ_x is precomputed from a large data base of LPC vectors. Hence, the decomposition specified in Eq (32) is precomputed saving the effort of decomposing the covariance matrix for every new LPC analysis frame. In fact, no improvement has been achieved when we tried to update the covariance matrix every LPC analysis frame.
Low spectral deviations were achieved when this method was used to quantise the LAR parameters with 25 bits per LPC update frame. A two-stage VQ approach was adopted to reduce the complexity of the error minimisation procedure. Exploiting the high correlation between the LSFs in adjacent frames, the method has given better results when the vector x to be quantised was the difference between the present LSF vector and the previously quantised one. Using two codebooks with 256 entries each and using 2 bits to quantise each codebook gain, the LPC prameters are quantised with a total of 20 bits, while maintaining good subjective speech quality. Curves of spectral deviation (SD) versus bit number per LPC analysis frame for five different LPC quantisation methods are depicted in Fig. 3. As bench markers, we have displayed the SD curves for the scalar quantisation of the LAR and LSF filter parameters, where the respective SD curves are denoted by SQ-LAR and SQ-LSF. Also shown are two VQ SD curves simulated according to [6] and [8], which are denoted by VQ2-LAR and VQ1-LSF, respectively. Finally, the SD curve representing our proposed method is denoted by VQ3-LSF. Observe that our results are close to those of the method suggested in [6] with the advantage of more robustness against channel errors. We are currently working on improving the quantiser performance by deploying a switched-adaptive approach [7], where a few number of fixed covariance matrices are used for different classes of speech.

4 Results and Discussion

In our experiments, we have investigated the performance of the three SEV Algorithms described in Section 2. We quantised the SEV parameters using different bit allocation

schemes resulting into bit rates of 3.2, 4.8 and 6.8 Kb/s for the three different Algorithms. We show here in Table 1 as an example the bit allocation of the 4.8 Kb/s coder using Algorithm 1. In Table 1, the LPC update frame is 25 ms, which corresponds to 200 samples of speech. The LPC frame is divided into five sub-blocks of 40 samples, and the LTP and excitation parameters are determined for each sub-block. A decimation factor D=4 is used resulting into decimated innovation sequences of length 10 only with a delay range of 64. This illustrates the simplicity of the excitation search algorithm. The LPC parameters are quantised using our proposed VQ scheme, where the vector to be quantised is given by the difference between the present and previously quantised LSF vectors. In Table 1, a total of 23 bits are allocated for the LPC parameters, where a two stage VQ scheme is used. The address lengths of the first and second codebooks are 10 and 9 respectively, and their gains are quantised using 2 bits each. Using 23 bits resulted into a spectral deviation below 0.8 dB, and no significant perceptual degradation was noticed compared to the unquantised case. The 4.8 Kb/s coder using Algorithm 1 has resulted into good communications speech quality according to our informal listening tests. The SEG-SNR at 4.8 Kb/s was 9.36 dB. At 6.8 Kb/s encoding rate, near-toll quality speech was achived with a SEG-SNR of 11.7 dB. When the bit rate was reduced to 3.2 Kb/s, the speech quality was rather synthetic.
There was a noticeable quality impairment when the speech produced by Algorithm 2 and Algorithm 3 was compared to the speech encoded by Algorithm 1 at the same bit rate. However, Algorithm 2 and Algorithm 3 are computationally more efficient than Algorithm 1, and at 4.8 Kb/s near communications quality has been achieved with dramatically reduced computational complexity.

5 Conclusion

By using the proposed stochastic VQ scheme for quantising the LPC parameters, we were able to efficiently encode the short-term spectral envelope of speech with a number of bits below 23, which did not result into serious spectral distortion. Deploying this VQ scheme in the suggested SEV resulted in communications quality speech at 4.8 Kb/s with dramatically reduced coder complexity.

Acknowledgement

The authors wish to express their sincere gratitude to prof. Raymond Steele for his continous encouragement and support.

REFERENCES

[1] R.C.Rose and T.P.Barnwell III, "Quality comparison of low complexity 4800 bps self excited and code excited vocoders," Proc. ICASSP' 87, pp. 1637-1640.
[2] N. Gouvianakis and C. Xydeas, "Advances in analysis by synthesis LPC speech coders," J. IERE, Vol. 57, No. 6 (supplement), pp. S272-S286, Nov/Dec 1987.
[3] M.R.Schroeder and B.S.Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," Proc. ICASSP' 85, pp. 937-940.
[4] J.Makhoul, S.Roucos ans H.Gish, "Vector quantization in speech coding," Proc. IEEE, Vol. 73, No. 11, pp. 1551-1588, Nov. 1985.
[5] Y. Shoham, "Vector predictive quantization of the spectral parameters for low rate speech coding," Proc. ICASSP' 87, pp. 2181-2184.
[6] K.Y.Lee, A.M.Kondoz and B.G.Evans, "Speaker adaptive vector quantisation of LPC parameters of speech," Electronics Letters, Vol. 24, No. 22, pp. 1392-1393, 27th Oct. 1988.
[7] M.Yong, G.Davidson and A.Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," Proc. ICASSP' 88, pp. 402-405.
[8] B.S.Atal, "Stochastic Gaussian model for low-bit rate coding of LPC area parameters," Proc. ICASSP' 87, pp. 2404-2407.

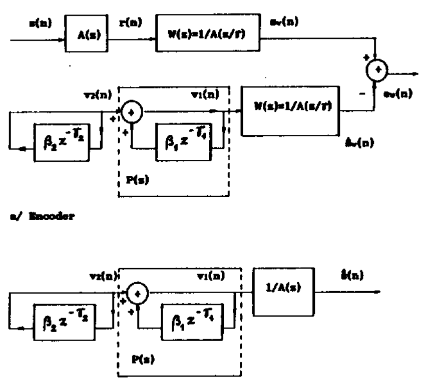

Fig. 1: SEV schematic diagram.

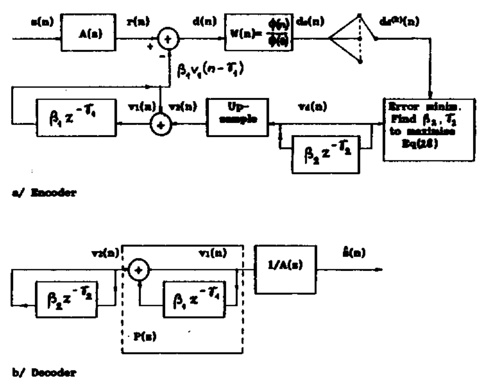

Fig. 2: RPE-SEV schematic diagram.

Parameter	No of bits
Short-term predictor	23
5 LTP delays	27
5 LTP gains	20
5 SELF loop delays	30
5 SELF loop gains	20
Total	120

Table 1: Bit allocation in a 25 ms frame for 4.8 Kb/s SEV using Algorithm 1.

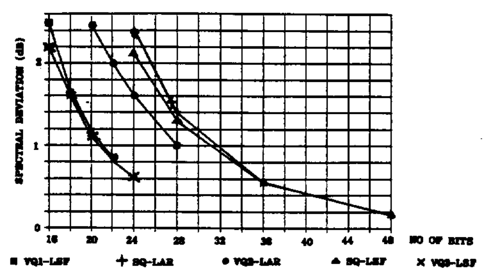

Fig. 3: Spectral distortion for different quantisation methods of LPC parameters.