Finite state CELP for variable rate speech coding

S.V. Vaseghi, PhD

Indexing terms: Speech synthesis, Coding, Predictive techniques

Abstract: The performance of a variable rate code excited linear predictor system is investigated. The coding system is based on a finite state CELP (FSCELP) frame work. Each individual state is primarily identified with a LPC model order, LPC coefficients bit allocation, excitation code book population density and state encoding rate. Successive input speech vectors are encoded at a rate that depends on the current state of the FSCELP system and the input vector characteristics. The use of a finite state system involves implicit clustering of speech signals. The lower rate states are selected during highly correlated steady state speech segments when relatively few bits are required to obtain adequate fidelity. For speech signals with a strong glottal excitation, unvoiced signals and transient speech segments, a relatively greater quantisation accuracy is needed to obtain good fidelity and therefore higher rate states of the system are used. Further improvement is obtained by using gamma populated excitation codebooks, for those states that are mainly used to encode speech signals with a strong underlying glottal excitation pulses. Experiments focus on investigation of the varying encoding requirements of the excitation signal for low pass, voiced, unvoiced and transient speech signals. The parameters of the finite state CELP system are designed to match the encoding requirements of typical speech signals. The greater part of the coding gain is obtained from variable rate encoding of the excitation signal. Using a six-state FSCELP, good quality speech is obtained at an average, maximum and minimum bit rates of 4 kbit/s, 10 kbit/s and 2 kbit/s, respectively.

1 Introduction

In speech communication, and storage systems for efficient utilisation of the available frequency and storage resources, it is desirable to compress the signal as much as possible while still retaining the intelligibility and a reasonable level of subjective quality. Compression in speech signals is achieved by taking advantage of the correlation and the redundancy in successive samples and frames of speech waveforms. In a fixed rate coder operating on blocks of speech samples the quantization scheme is based on the measured average contribution of speech parameters to the perceptual quality and intelligibility of the synthesised speech signals.

Speech production and perception are highly nonstationary processes. The quantisation accuracy (source coding rate) that is required to reproduce a speech segment at a given level of fidelity varies considerably with the speech spectral characteristics. Some studies have focussed on the effects of dynamic bit allocation to speech parameters in fixed rate systems. In these systems the frame encoding rate is fixed but the distribution of bits within a frame is variable and depends on the short term characteristics of speech signals. In Reference 1 a dynamic bit allocation sub-band coder has been described. Jayant and Chan [2] have reported on some experiments on the dynamic allocation of bits to LPC parameters and the excitation signal of a code excited linear prediction system operating at several different rates. Exhaustive search methods are used to find the best bit allocation pattern out of several LPC parameter-excitation signal quantisation patterns. They obtained an average improvement of 1–2 dB and higher gains for certain speech segments. However it is generally believed that the small improvement which results from the dynamic bit allocation methods is not perceptually significant.

Complete utilisation of the time varying characteristics of speech signal is only possible with the use of a variable rate coding system. Generally a variable rate coding system matched to the time varying characteristics of the source achieves higher SQNR than a fixed rate system operating at the average coding rate. Variable rate coding systems are particularly useful for voice response systems where computations can be performed off-line and the delay is not a problem. Computational delay and the buffer delay place some constraints on the use of a variable rate coding system in a real time communications systems. The main trade off in source coding is between the rate and distortion. In varying rate systems the bit rate is increased for those speech segments that require finer quantisation for adequate fidelity, such as transients and unvoiced speech, and is decreased for highly correlated steady state segments that can be represented with a fewer number of bits. The coding rate is normally obtained from a rate-distortion cost function that reflects the desire to have good quality (small distortion) speech at low bit rate.

Most of the early work on variable rate coding has focussed on the time varying characteristics of the correlation structure in the LPC parameter vectors where new speech parameters are transmitted when the change from the previously transmitted parameters makes a significant contribution to the speech quality given the constraints on the transmission rate. Magill [3] proposed a system in which a new reflection coefficient vector was transmitted only when it showed a significant change from the previous value. Makhoul extended the work and reconstructed the missing parameters by linear interpolation between the transmitted frames [4]. Papamichalis and Barnwall proposed a system in which the alternatives were to transmit all, none or a subset of the reflection

Paper 8272I (E5, E7, E8), first received 4th July 1990 and in revised form 30th May 1991

The author is with the School of Information Systems, University of East Anglia, Norwich NR4 7TJ, United Kingdom

coefficients [5]. Their basic LPC system uses a 10th order model. To keep the computation costs down to implementable level only subsets of four or eight coefficients are transmitted. Dynamic programming was used to determine the optimum choice of parameters from 16 consecutive frames of speech signal. The cost function penalised for increasing bit rate and signal distortion. These systems are based on the observation that the quantisation accuracy required to represent speech spectral information parameters at a given level of fidelity varies with the time varying characteristics of the speech signals. These variable rate systems were aimed at very low bit rate vocoders where a significant fraction of the bit resources are used to encode the LPC filter parameters and only minimal information (voicing, pitch period) about the excitation signals are transmitted. Consequently they did not address the problem of variable rate encoding of the excitation signal which is of primary importance in the design of good quality speech coders.

Ayanoglu and Gray [6] used the LPC model of speech and predictive trellis waveform coding to find the optimum excitation pattern from an excitation codebook. The excitation codebook consisted of scalar quantised values which were obtained from either a Gaussian distribution or clustering of a training speech data set. In a system described in Reference 7 speech signals are segmented into phonetically distinct categories. Each segment type is labelled and encoded using a bit allocation pattern that is based on the speech characteristics within the segment and the contribution of the different speech parameters to perceptually important features of the segment. In addition to the use of variable signal characteristics the entropy of the excitation code book entries can be used to achieve further compression. An iterative descent algorithm based on a Lagrangian formulation was introduced by Chou *et al.* [8] for designing optimum vector quantisers subject to an entropy constraint.

This work focuses on the design and performance of a variable rate code excited linear predictor based on a finite state vector quantisation (or a finite state CELP FSCELP) frame work. The states of the system are primarily identified with the state coding rate and state code book population and each state uses a distinct combination of the LPC coefficients and excitation quantisation pattern that is specific to that state. Low bit rate states use smaller code books for coding. These states are mainly used to encode highly correlated low frequency speech segments. Higher rate states are used mainly during transients and speech segments with significant wide band components. As in almost all CELP coders the LPC parameters are scalar quantised to keep the design efforts to a minimum and ensure performance robustness.

2 Code excited linear prediction

Speech signals are commonly modelled as the output of a linear prediction system excited by a wide band excitation. The speech model is

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} + Ge_n \tag{1}$$

or in the z-domain

$$P_1(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

where $\{x_n\}$ are speech samples, $\boldsymbol{a}$ is the LPC parameter vector which represents the vocal tract system, $\{e_n\}$ is the excitation sequence and the scalar G is the gain of the linear predictor. The long term correlation in the excitation signal which is primarily attributed to the pitch periodicities are modelled by a pitch prediction filter β

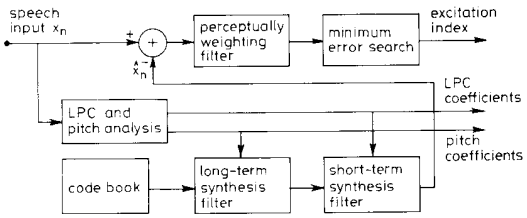

Fig. 1 *Code-excited linear prediction*

$$P_2(z) = \sum_{k=q}^{q} \beta_k z^{-m-k} \tag{3}$$

Normally the pitch filter $P_2(z)$ is implemented as a filter of one to three taps long. The coefficients of the so called short term prediction filter $\boldsymbol{a}$ are obtained from speech data using the least square error formulation. The long term correlation lag m in eqn. 3 is obtained from the location of the peak of the autocorrelation of the speech signal or that of the excitation signal in a prespecified range. The optimal coefficients of the pitch prediction filter are calculated from the excitation signal using the least square error formulation. To summarise, in this model speech is decomposed into three basic constituents:

(*a*) short term prediction filter $P_1(z)$
(*b*) pitch prediction filter $P_2(z)$
(*c*) white excitation sequence e_n and gain g

Among the various methods of respresentation of the LPC vector $\boldsymbol{a}$, the PARCOR coefficients, the log area ratio and the line spectrum pairs are widely used for their robustness and good quantisation properties. The speech coder is based on LSP speech analysis-synthesis [9]. The frequency structure (frequency ordering property) and time correlations (that is the interframe and intraframe) structure of the LSP parameters can be used to produce efficient LSP quantisers. The pitch prediction filter coefficients are scalar quantised and transmitted. In high quality speech encoding systems, representation of the excitation signal consumes the greater part (up to 75%) of the available bit resources and vector quantisation methods is used to compress the excitation.

Code excited linear prediction (CELP) is an efficient closed loop analysis-synthesis method for narrow and medium band (4–16 kbit/s) speech coding systems, Fig. 1 [10–12]. In CELP coders speech is segmented into frames which are typically 10–30 ms long. For each segment an optimum set of linear prediction and pitch filter parameters are determined and quantised. Each speech frame is further subdivided into a number of subframes of equal length (typically 5 ms). For each subframe an excitation codebook is searched to find the input vector to the quantised predictor that gives the minimum mean squared error reproduction of the speech signal. The index of the vector quantised excitation signal and the scalar quantised values of LPC, pitch filter and gain parameters are channel encoded and transmitted. The closed loop analysis, i.e. the determination of the input reproduction vector that produces the minimum distortion synthesised speech, is the feature that makes

CELP more powerful than the traditional open loop voice coding systems.

CELP coders operating at low bit rates are not able to reproduce high frequencies or rapid transients in speech signals. Below one bit per sample the reconstructed signal suffers from a degradation that is more noticeable for high pitched voices with strong glottal excitation pulses such as female voices [12]. The problem is partly caused by the limited size of the excitation code book and its population density and partly caused by fixed frame length LPC analysis where one set of LPC parameters is used to describe spectral information about a frame of speech that may obtain widely different speech events. In general the reproduction accuracy improves with increasing excitation code book size and when the code book population is a better approximation to the speech excitation distribution. For instance during voiced speech segments with strong non-Gaussian distribution the use of Laplacian or gamma code book produces better result than a Gaussian code book. The segmental SQNR also improves when speech frames are chosen such that they contain acoustically homogeneous segments.

Conventionally in CELP coders, only the excitation signal is vector quantised. The LPC parameters may also be vector quantised. However, for simplicity of design and performance robustness, the encoding of the LPC parameters is commonly achieved by the scalar quantisation of the equivalent set of reflection coefficient, log area ratios or line spectrum pairs. More recently differential line spectrum pairs frequencies are used because of their robust quantisation properties in the design of voice coding systems [13].

In code excited linear prediction for each input vector $\boldsymbol{x}_n$, every code book vector is tested to find the best reproduction vector $\hat{\boldsymbol{x}}_n$:

$$\hat{x}_n = \sum_{k=1}^{p} a_k \hat{x}_{n-k} + Ge_n \tag{4}$$

The signal distortion or speech waveform quantisation noise is given by

$$d_n = x_n - \hat{x}_n = \sum_{k=1}^{p} a_k d_{n-k} + q_n \tag{5}$$

where $q_n = e_n - \hat{e}_n$ is the excitation quantisation noise. Note that in CELP, the coder noise is the output of the LPC filter, $\boldsymbol{a}$, driven by the excitation quantisation noise q_n. That is the excitation quantisation noise is spectrally shaped and amplified by the LPC filter. This is unlike ADPCM systems in which the predictor is placed inside a feedback loop such that the speech waveform quantisation noise is equal to the excitation (residual) quantisation noise. In a CELP system assuming that q_n is white, the quantisation noise will have the same spectrum as the LPC model of speech $\boldsymbol{a}$. It can be easily shown that the SQNR of the reconstructed speech is equal to SQNR of the excitation reproduction signal.

To improve the quality of synthesised speech a perceptually weighting filter can be applied to the input speech Fig. 1. Commonly this filter has the general form

$$W(z) = \frac{1 - \sum_{i=1}^{p} \alpha^i a_i z_i}{1 - \sum_{i=1}^{p} \beta^i a_i z^i} \quad 0 \leqslant \alpha, \beta \leqslant 1 \tag{6}$$

In the case when $\beta = 0$ and $\alpha = 1$, $W(z)$ is the inverse LPC filter and the search method finds the codebook entry which is the best reproduction of the excitation signal. On the other hand when $\alpha = \beta = 1$ (or $\alpha = \beta = 0$), $W(z) = 1$, and the search method finds the codebook entry that produces the best reproduction of the speech signal itself [12].

In conventional CELP coders up to 75% of the available bit resources are used for representation of the excitation signal. However the excitation quantisation accuracy that is required to reproduce a speech segment with a desired SQNR is strongly dependent on the characteristics of the speech segment. Lowpass highly correlated speech segments can be reproduced with relatively coarse representation of the excitation signal. On the other hand more bits are needed to encode the excitation during unvoiced segments, transients and voiced segments with strong glottal excitation. The excitation encoding rate necessary to achieve a given level of fidelity is time varying and this is where the major savings can be obtained. Similarly, variable rate compression in encoding of the LPC parameters can be achieved by using different LPC model order and bit allocations, for different types of speech signals.

3 Finite state code excited linear predictor

A finite state CELP system for variable rate speech coding is shown in Fig. 2. It is essentially a collection of CELP vector quantisers operating in parallel. The state parameters of a FSCELP are the LPC model order, LPC coefficients bit allocation, excitation population density and state encoding rate. The state parameters of an N state system are designed to match the encoding requirement of N distinct classes of speech signals. Each input speech vector is encoded using a state that achieves a good reproduction at a minimum bit rate.

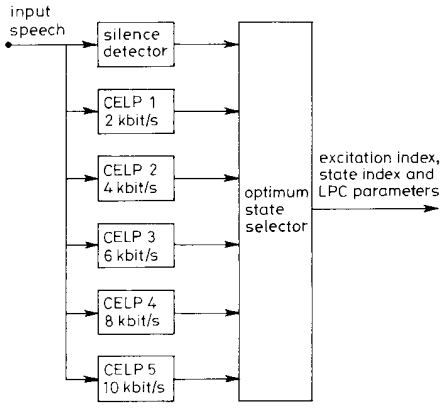

Fig. 2 *Six-state finite state CELP coder*

In general a finite state vector quantiser (FSVQ) is defined by the number of states N, the set of state codebooks $C_s = \{C_1, C_2, \ldots, C_N\}$, state coding rate $\{R_1, R_2, \ldots, R_N\}$ and state transition rule $\{T_{ij}\}$ [6]. The state transition rule of a FSVQ can be either deterministic or based on a rate distortion cost function. The next state map of the FSVQ can be obtained from a training data set such that the current state and the code word selected determine the next encoder state [6]. At the receiver the entire state sequence is obtained from the initial state and the transmitted codewords. Such a deterministic next map requires considerable training of the system across a

wide range of speaking population. The aim of this work is to investigate the performance of a finite state speech compression system that is both robust and relatively simple to implement. For this reason the transition rule chosen is not deterministic but is a rate-distortion function of the current state and the input vector.

The main advantage of a variable rate coding system is the ability to increase the encoding rate and avoid excessive distortion of those speech segments that require a finer quantisation for good fidelity. So long as the signal characteristic does not change significantly the current state is expected to remain occupied. A transition to a new state of the machine indicates a change in the speech characteristics such that a different rate-distortion balance is required. It may therefore be expected that the system moves to a given state during speech segments of a certain characteristic. For example highly correlated lowpass sections of speech will cause the system to move to a lower rate state and conversely during an unvoiced passage with a relatively white frequency spectrum the quantiser will move to a higher encoding rate state. In a way an N-state FSVQ partitions speech into N sets of different characteristics. For simplicity of the system design procedure and operational robustness the FSVQ considered is based on the scalar quantisation of the LPC parameters.

In finite state vector quantisation a cost which is an increasing function of the state encoding rate and the distortion is assigned to each state of the system. The cost function may include the effect of each codebook entry on the future as well as the present performance of the system. In delayed decision systems the consequence of selecting an state on future performance of the system is considered and at each stage an encoding system is selected that minimises a cumulative cost function. The cost function is designed such that it reflects the desire to achieve good fidelity at low bit rates and it penalises for increasing bit rate and distortion. The transmission cost at time n may be expressed as

$$C(D, R, n) = R(n) + \alpha_n D(n) \tag{7}$$

where $R(n)$ is the source coding rate and $D(n)$ is the quantisation distortion. The scalar parameter α_n is a control parameter that can be used to operate the coder at an average coding rate within the bounds of a predetermined maximum and minimum bit rate. The cost of transmission at the rate R is simply the number of bits used to encode an input vector [5].

$$R(n) = b(n) \tag{8}$$

The distortion part of the cost function should correlate well with the perceptual quality of speech signals. A weighted mean squared error is commonly used to express the quantisation distortion

$$D(n) = 10 \log_{10}(\sum w_n(x_n - \hat{x}_n)^2) \tag{9}$$

At each time instance, n, the state selected is that which minimises cost $C(D, R, n)$.

4 Experimental results

The speech data set used in the following experiments consisted of 10 mins of a telephone conversation of a male and a female speaker sampled at a rate of 8 kHz and quantised to 16 bit precision. The speech data set was supplied by the British Telecom Research Laboratories (Martlesham Heath). Speech were segmented into frames of 32 ms (256 sample) long. The LPC parameters for each frame were obtained by the Burg's method [14].

The LPC model order, LPC coefficients bit allocation, and in particular the number of bits per excitation sample that is required to reproduce a speech segment with a given level of fidelity, strongly depends on the speech signal characteristics. The results of experiments on encoding some typical speech segments by CELP coders are described. These experiments illustrate the variable rate encoding requirement of the excitation signals. Fig. 3*a* is a speech segment from a male speaker with a dominantly low pass characteristics. The signal is highly resonant and can be well reproduced with a relatively coarse quantisation of the excitation signal. For this speech segment using a 6th order LPC model, an excitation signal of 0.125 bits per sample achieved 14 dB SQNR. The synthesised signal is shown in Fig. 3*b*.

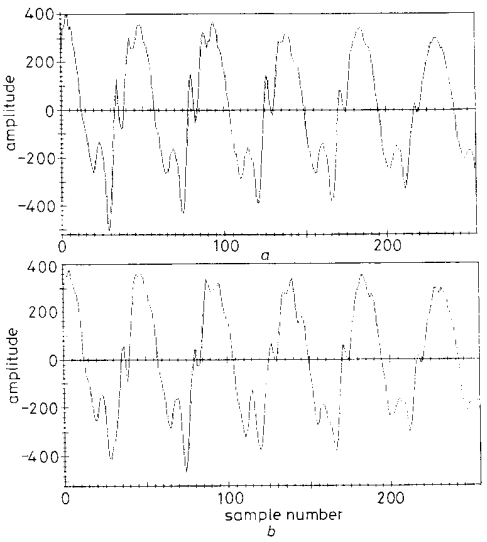

Fig. 3 *Male speech with dominantly lowpass and resonant characteristics*

a Typical sample
b Synthesised speech (0.125 bits per excitation sample; 6th order predictor; SQNR = 14 dB)

Fig. 4*a* is a voiced speech segment of a female speaker with strong glottal excitation. The excitation signal which contains quasi-periodic pulses is non-Gaussian. A reproduction of this signal using a Gaussian code book at an encoding rate of 1 bit per excitation sample is shown in Fig. 4*b*. This Figure demonstrates that the Gaussian population density commonly used for CELP excitation codebook are not suitable for voiced speech segments with strong glottal excitation pulses. The glottal excitation pulses can have amplitudes which are much larger than the short time mean of the excitation signal. These large outliers can not be reproduced from a Gaussian code book. An alternative to the use of a Gaussian density is a gamma density populated code book. Fig. 4*c* shows the reproduction of the signal using a gamma populated code book. In this case the use of a gamma code book resulted in an improvement of 3.6 dB. Another alternative is to use a code book populated with actual speech excitation obtained from speech residuals. Clustering methods such as the K-means algorithm may be used to obtain an optimal code book from a large data

set of speech residuals. In any case the use of a non-Gaussian excitation for synthesis of strongly voiced signals may improve the reproduction accuracy.

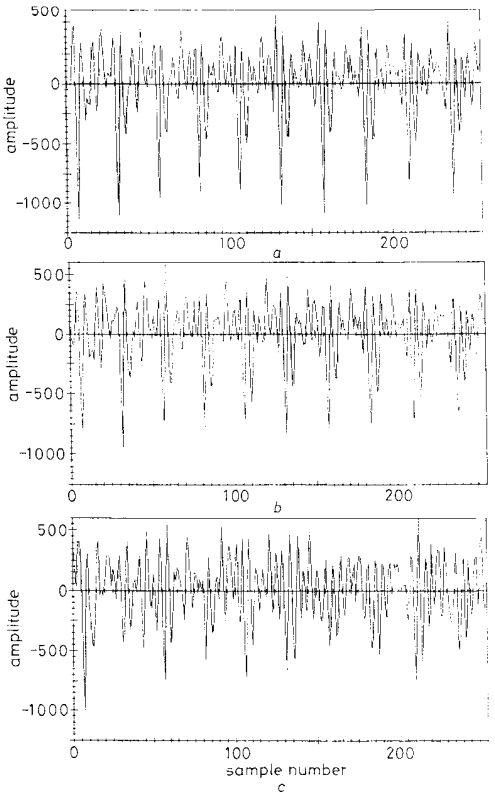

Fig. 4 *Female speech with strong glottal excitation pulses*
a Typical sample
b Synthesised speech (1 bit per excitation sample; 12th order predictor; SQNR = 7.4 dB) Gamma code book
c Synthesised speech (1 bit per excitation sample; SQNR = 12 dB) Gaussian code book

Fig. 5*a* is a section of unvoiced speech signal. This noise-like signal can be adequately modelled by a LPC system of four to six poles. However finer quantisation of the excitation signal is required for good fidelity. For this signal using a 4th order model an excitation signal of 1 bit/s achieved 8 dB SQNR (Fig. 5*b*). In general more bits/sample are required to reproduce transient signals, signals with strong excitation pulses and signals with a relatively wideband frequency spectrum.

Fig. 6 shows the variation of average SQNR with the number of bits per excitation sample for unvoiced, strongly voiced, voiced and lowpass speech signals. The lowpass type classification refers to speech signals which have a predominantly low frequency composition. The strongly voiced type classification refers to speech which exhibits strong glottal excitation pulses. For each class of speech signal the SQNR statistics were obtained from 200 manually segmented speech signals which were hand picked from a relatively large data base. Each speech segment was 256 samples long. The Figure demonstrates that the SQNR achieved at a given number of bits/excitation sample depends strongly on the speech signal. Finer quantisation are needed for signals with a relatively wide-band frequency spectrum such as unvoiced signals or signals which have strong glottal excitations. The plot shows that for unvoiced or strongly voiced signal up to 1 bit per excitation sample is required to achieve up to 8 dB SQNR. On the other hand voiced and lowpass speech segments may be encoded at below 0.5 bits per excitation sample.

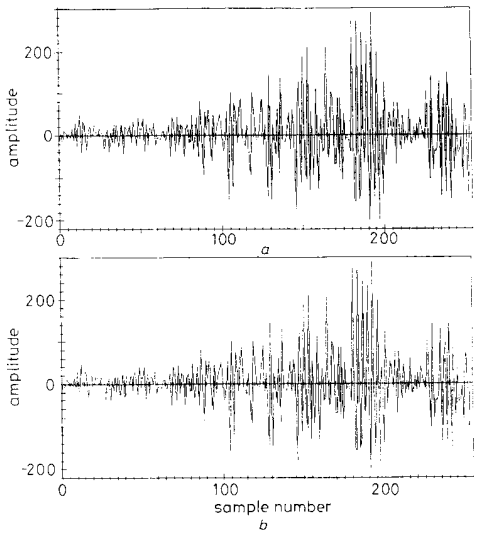

Fig. 5 *Unvoiced male speech*
a Typical segment
b Synthesised speech (1 bit per excitation sample; 4th order predictor; SQNR = 8.5 dB)

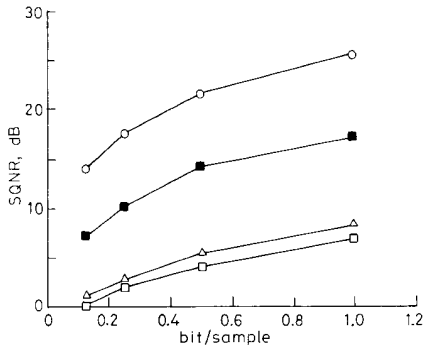

Fig. 6 *Variation of SQNR with bits per sample*
□ unvoiced
△ strongly voiced
■ voiced
○ lowpass

The gain in bits per sample of a variable rate coding scheme (compared with a fixed rate scheme) depends on the average proportion of each speech type in a typical conversation. Fig. 7 shows a histogram of various speech types in a data set of 5 min conversation of a male and a female speaker. Speech signals were classified into five classes of silent, lowpass, voiced, strongly voiced and unvoiced. Long silent events were excluded and only short silent pauses of less than 3 s were included. From the histogram about 30% of speech signal consists of short pauses. Low pass speech signals requiring less than 0.2 bit/sample constituted about 20% of the conversa-

tion. Voiced signals constituted 28% of speech population. Voiced signals may need about 0.2 to 0.5 bit/sample for adequate fidelity. The population of highly voiced signals, with strong underlying glottal excitation, and unvoiced speech were about 12% and 10%, respectively. This is fortunate as these signals require the most amount of bits per sample for adequate reproduction.

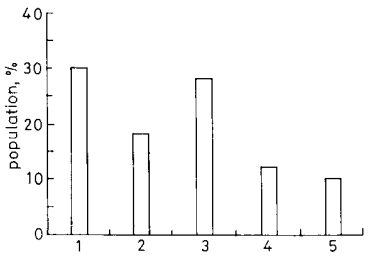

Fig. 7 *Percentage population of silence (1), lowpass (2), voiced (3), strongly voiced (4) and unvoiced (5) signals in 5 min telephone conversation*

Table 1 is a state bit allocation for a five state CELP with an additional state for encoding of silent periods. With six states 3 bits are needed to signal the encoding state used at the transmitter. To reduce the overhead due to this side-information, a minimum state occupancy duration can be imposed. If each state when visited remains occupied for a minimum of one frame duration which is 32 ms then 94 bit/s of side information are needed to signal the state information.

Of interest are the time varying segmental SQNR achieved by each state CELP when it is employed to encode the entire speech data set by itself in the fixed rate mode. Fig. 8 shows the SQNR histograms for each of the 5 CELP coders listed in Table 1. Each histogram was obtained by using one CELP to encode the entire speech data set segmented to frames of 256 samples. The histograms show the percentage of frames (256 samples long) that were encoded by each CELP at various SQNR. For example from Fig. 8*a* it can be seen that about 15% of the speech data set can be encoded at 0.125 bit/excitation sample with a SQNR of more than 10 dB. At this encoding rate more than 45% of frames will be encoded at less than 2 dB SQNR. Each histogram indicates the percentage of the input signal that can be encoded with a desired degree of fidelity at a given bit rate.

A six-state FSCELP implicitly classifies the input signal into one of six classes. In Table 1, state 1 is the speech inactive state. During silence periods a very low level noise is output at the receiver to induce a natural sensation. State 2 is primarily designed for the use by lowpass speech events. These signals can be adequately modelled by a 6th order linear predictor and a relatively coarse quantisation of the excitation. In this state 1 kbit/s are allocated to the LPC parameters and 0.125 bit/sample (1 kbit/s) to the excitation signal. States 3, 4 and 6 each use a 12th order LPC model and have 1.6 kbit/s allocated to LPC parameter quantisation. State 3 and 4 allocate 0.25 bit/sample (2 kbit/s) and 0.5 bit/sample (4 kbit/s) to excitation quantisation, respectively. These states are designed for use by various types of voiced signals. State 6 has 1 bit/sample (8 kbit/s) allocated to the excitation signal quantisation. For this state a gamma populated code book is used. This state is primarily intended for speech signals which have very strong glottal excitation pulses. State 5 is intended for encoding of

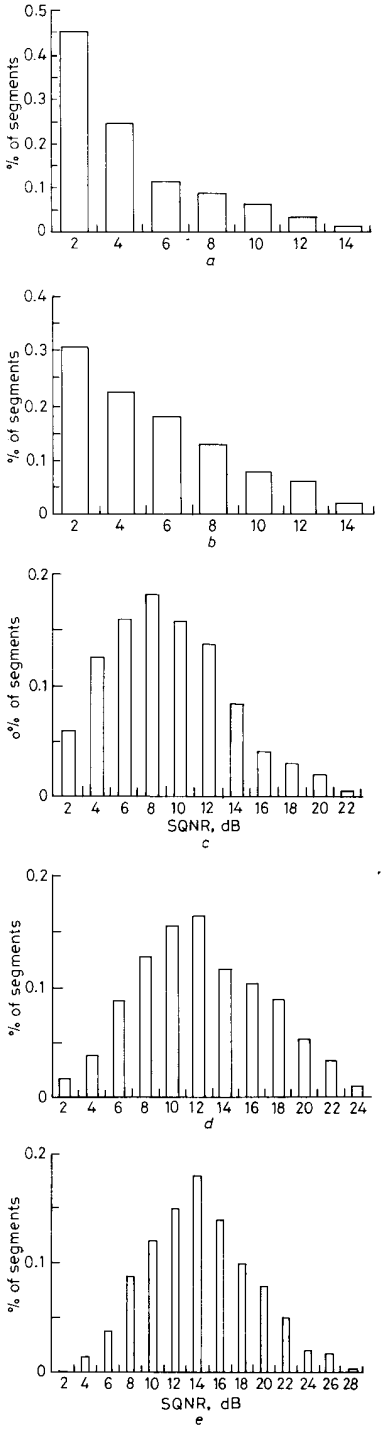

Fig. 8 *Histogram of SQNR for CELP coders*

a 0.125 bit/excitation sample
b 0.25 bit/excitation sample
c 0.5 bit/excitation sample
d 0.75 bit/excitation sample
e 1 bit/excitation sample

unvoiced speech. For this reason a 4th order model is used and 600 bit/second are allocated to LPC parameter quantisation. For excitation quantisation state 5 uses 1 bit/sample.

number of classes. The parameters of each state are the LPC model order, LPC coefficients bit allocation, excitation population density and the bit rate. The state parameters are designed to match the encoding require-

Table 1: State bit allocation for five-state variable rate CELP code

State overall, bit rate, bit/s	LPC parameters				Excitation parameters			
	frame size, ms	model order	bits per frame	bits per second	sub-frame size, ms	bits per subframe	bits per sample	bit/s
Silence	—	—	—	—	—	—	—	—
2000	32	6	32	1000	8	8	0.125	1000
3600	32	12	52	1600	4	8	0.25	2000
5600	32	12	52	1600	2	8	0.5	4000
8600	32	4	20	600	1	8	1.0	8000
9600	32	12	52	1600	1	8	1.0	8000

4.1 FSCELP Operating a minimum SQNR constraint

In this experiment for each input speech vector the encoding state selected is the minimum rate state that achieves a SQNR above a predetermined minimum level. This encoding rule that achieves maximum compression subject to a minimum SQNR constraint gives simple control of distortion level and is particularly useful for voice response systems and applications in which distortion control is the major consideration. A computational advantage of the method is that low rate states are searched first and when the minimum rate state that achieves the desired SQNR is found the search for higher rate states are avoided. Table 2 lists the system parameters of interest for a minimum SQNR constraint of 10 dB.

Table 2: Finite state CELP parameters operating at 10 dB minimum SQNR constraint

State	Occupancy rate, %	SQNR, dB		
		average	minimum	maximum
1	0.13	14.5	10.4	20
2	0.2	13.5	10.0	18.5
3	0.22	12.0	10.7	16.5
4	0.1	11.0	7.0	13.0
5	0.08	12.0	8.0	14.0
Silence	0.27	—	—	—

It should be pointed out that depending on the minimum SQNR desired and the maximum encoding rate available, the SQNR of some speech segments may fall below the desired level even for the highest encoding rate state. At a minimum distortion constraint of 10 dB the average encoding rate is 3.8 kbit/s with an average SQNR of 12.4 dB. The fraction of the input signal segments for which the maximum rate state was unable to achieve the minimum 10 dB SQNR was 0.05. The quality of the variable rate encoded speech at an average rate of 4 kbit/s was judged as good as an 8 kbit/s fixed rate CELP coder. For use in a real communication system the computational delay and the buffer delay makes a variable rate system less attractive. A more attractive and growing application area is the automatic voice response systems, in which the computations can be performed offline.

5 Conclusion

A finite state CELP coder for variable rate speech coding has been presented. It was pointed out that this system implicitly clusters the input speech signal into one of a ments of the most frequently occurring and distinct types of speech. For the state which is expected to be employed for encoding of voiced signals, with strong underlying excitation pulses, a gamma populated excitation code book is used. Each input signal is encoded using a state that achieves the minimum encoding rate subject to a desired fidelity constraint. Some results are presented that describe the encoding requirements and the frequency of occurrence of lowpass, voiced and unvoiced speech signals. Experiments show that the finite state CELP at an average rate of 4 kbit/s, with maximum and minimum rates of 10 kbit/s and 2 kbit/s, sounds as good as an 8 kbit/s fixed rate CELP coder. The system can be improved by using a greater number of states whose state parameters are chosen such that they closely match the encoding requirements of typical speech events.

6 Acknowledgment

The author would like to thank Dr. Ivan Boyd and acknowledge the support of the British Telecom Research Laboratory, Martlesham Heath, Ipswich, UK.

7 References

1 COX, R.V., GAY, S.L., SHOHAM, Y., QUACKENBUSH, S.R., SESHADRI, N., and JAYANT, N.S.: 'New directions in subband coding', *IEEE J. Sel. Areas Commun.*, 1988, **6**, (2), pp. 391–409
2 JAYANT, N.S., and CHEN, J.H.: 'Speech coding with time-varying bit allocations to excitation and LPC parameters'. IEEE Proc. ICASSP 89, **1**, May 1989, Glasgow, pp. 65–68
3 MAGILL, D.T.: 'Adaptive speech compression system for packet communication systems'. Telecomm. Conf. Record, 1973
4 MAKHOUL, J., VISWANATHAN, R., COSSELL, L., and RUSSEL, W.: 'Natural communication with computers: speech compression at BBN'. BBN Report No. 2976, Vol. 2, 1974
5 PAPAMICHALIS, P.E., and BARNWELL, T.P.: 'Variable rate speech compression by encoding subsets of the PARCOR coefficients', *IEEE Trans.*, 1983, **ASSP-31**, (3), pp. 706–713
6 AYANOGLU, E., and GRAY, R.M.: 'The design of predictive trellis waveform coders using the generalised Lloyd algorithm', *IEEE Trans.*, 1986, **COM-34**, (11), pp. 1073–1080
7 WANG, S., and GERSHO, A.: 'Phonetically-based vector excitation coding of speech at 3.6 kbps'. IEEE Proc. ICASSP 89, **1**, May 1989, Glasgow, pp. 49–52
8 CHOU, P.A., LOOKABAUGH, T., and GRAY, R.M.: 'Entropy constrained vector quantisation', *IEEE Trans.*, 1989, **ASSP-37**, (1), pp. 31–42
9 SUGAMURA, N., and ITAKURA, F.: 'Speech analysis and synthesis methods developed at ECL in NTT — from LPC to LSP', *Speech Commun.*, June 1986, **5**, pp. 199–215
10 ATAL, B.S., and SCHRODER, M.R.: 'Stochastic coding of speech at very low bit rates'. Proc. ICC, Amsterdam, May 1984, pp. 1610–1613

11 SCHRODER, M.R., and ATAL, B.S.: 'Code-excited linear prediction (CELP): high-quality speech at low bit rates'. Proc. IEEE-ICASSP, April 1985, pp. 937–940
12 KROONAND, P., and ATAL, B.S.: 'Strategies for improving the performance of CELP coders at low bit rates'. ICASSP 88, **1,** New York, April 1988, pp. 151–154
13 SUAMURA, N., and FARVARDIN, N.: 'Quantiser design in LSP speech analysis-synthesis', *IEEE J. Select. Areas Commun.*, 1988, **SAC-6,** (2), pp. 432–440
14 RABINER, L.R., and SHAFER, R.W.: 'Digital processing of speech signals' (Prentice-Hall, 1978)
15 SHANNON, C.E.: 'Coding theorems for a discrete source with a fidelity criterion'. IRE Nat. Conv. Rec., 1959, pp. 142–163
16 BURGER, T.: 'Rate distortion theory: a mathematical basis for data compression' (Prentice-Hall, Englewood, 1971)
17 YONG, M., and GERSHO, A.: 'Vector excitation coding with dynamic bit allocation'. Proc. IEEE-GLOBCOM, December 1988, pp. 290–294
18 JOHNSTON, J.D.: 'Transform coding of audio signals using perceptual noise criteria', *IEEE J. Sel. Areas Commun.*, 1988, **6,** (2), pp. 314–323
19 CUPERMAN, V.: 'On adaptive vector transform quantisation for speech coding', *IEEE Trans.*, 1989, **COM-37,** (3), pp. 261–267
20 ITAKURA, F.: 'Minimum prediction residual principle applied to speech recognition', *IEEE Trans.*, February 1975, **ASSP-23,** pp. 67–72
21 FOSTER, J., GRAY, R.M., and DUNHAM, M.O.: 'Finite-state vector quantisation for waveform coding', *IEEE Trans.*, May 1985, **IT-31,** pp. 348–359
22 STEWART, L.C., GRAY, R.M., and LINDE, Y.: 'The design of trellis waveform coders', *IEEE Trans.*, April 1982, **COM-30,** (4), pp. 702–710
23 BEI, C., and GRAY, R.M.: 'Simulation of vector trellis encoding systems', *IEEE Trans.*, 1986, **COM-34,** (3), pp. 214–218
24 EPHRAIM, Y., and GRAY, R.M.: 'A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantisation', *IEEE Trans.*, 1988, **IT-34,** (4), pp. 826–834
25 SOONG, F.K., and JUANG, B.H.: 'Line spectrum pair (LSP) and speech data compression'. IEEE ICASSP84, 1984, pp. 1.10.1–1.10.4