Speech Communication 10 (1991) 277–301
North-Holland

Performance study of stochastic speech coders

Bernt Ribbum, Andrew Perkis, K.K. Paliwal and Tor Ramstad
Division of Telecommunications and Acoustics, ELAB–RUNIT, The Norwegian Institute of Technology, Trondheim, Norway

Received 26 September 1990
Revised 30 April 1991

Abstract. Four speech coders in the CELP (Code Excited Linear Predictive) family are described. By replacing the long term prediction with a self excitation sequence (adaptive codebook) as well as substituting a stochastic multipulse, or sparse codebook for the commonly used codebook of Gaussian noise, the speech quality is improved. The coders are fully quantised at 7.0 and 5.0 kbit/s, interesting bit-rates for such possible applications as the half rate GSM system and the INMARSAT-M service. The performance of the coding schemes are evaluated by a formal listening test and presented by their Mean Opinion Scores (MOS). For the coder with maximum performance with respect to quality and complexity, a set of tolerable bit error rates (BER) are given. It is shown that the presence of acoustic background noise does not influence the coder quality, and that bit errors in this case will be partially masked by the background noise giving the coder a high degree of robustness. Considering the performance in the presence of bit errors and background noise, the coder seems to be suitable for use in a mobile communication system using satellite links.

Zusammenfassung. Vier Sprachkodierer werden beschrieben, die der CELP (Code Excited Linear Predictive; Kode-angeregter Kodierer mit linearer Voraussage) Familie angehören. Die Sprachqualität wird dadurch verbessert, daß die Langzeit (Tonfrequenz) Prediktion durch eine selbstanregende Sequenz (adaptative Kodierungstabelle) ersetzt wird, sowie durch Substituierung der häufig benutzten Kodierungstabelle nach Gauß durch einen stochastischen Mehrfachpuls bzw. eine schlichte Kodierungstabelle. Die Kodierer sind voll quantifiziert mit 7.0 und 5.0 kbit/s – eine interessante Übertratungsrate für mögliche Anwendungsgebiete wie das Halbgeschwindigkeits–GSM System und der INMARSAT-M Dienst. Die Leistungen der Kodierschemata werden durch einen formellen Hörtest evaluiert und durch die durchschnittlichen Meinungswerte wiedergegeben. Eine Reihe von annehmbaren Bit Fehlerraten wird für den Kodierer mit der besten Leistung in bezug auf Qualität und Komplxität angegeben. Es wird aufgezeigt, daß das Vorhandensein von akustischen Hintergrundgeräuschen die Kodierqualität nicht beeinflußt, und daß Bitfehler in diesem Fall teilweise durch die Hintergrundgeräusche verdeckt werden, wodurch der Kodierer einen hohen Grad an Robustheit gewinnt. Wenn die Leistung unter Berücksichtigung von Bitfehlern und Hintergrundgeräuschen evaluiert wird, scheint der Kodierer für den Gebrauch in einem mobilen Kommunikationssystem unter Anwendung von Satellitenverbindungen geeignet.

Résumé. Nous présentons dans cet article quatre codeurs de parole de la famille CELP (Code Excited Linear Predictive). La qualité de la parole est améliorée en remplaçant la prédiction à long terme par une séquence d'auto-excitation (dictionnaire de codage adaptatif), ainsi qu'en substituant un dictionnaire de codage réparti à l'ensemble des vecteurs codes à bruit gaussien couramment employé. Les codeurs sont entièrement quantisés à 5 et 7 kbit/s, c'est-à-dire à des vitesses de transmission intéressantes pour d'éventuelles applications telles que les systèmes GSM à demi débit et INMARSAT-M. Les performance de ces systèmes de codage sont évaluées grâce à un test conventionnel d'écoute et présentées par leur "Mean Opinion Scores (MOS)". Un ensemble de taux d'erreurs binaires tolérables est fourni pour le codeur dont les performances en fonction de la qualité et de la complexité sont maximales. Il est montré que la présence de bruit de fond acoustique ne modifie en rien la qualité du codeur et que dans ce cas les erreurs binaires seront partiellement masquées par le bruit de fond, grâce à la robustesse du codeur. Compte tenu des performances en présence de ces perturbations, le codeur semble approprié pour être utilisé dans des systèmes qui ont recours à des liaisons par satellite.

Keywords. Digital signal processing, speech coding, stochastic coders.

1. Introduction

A number of speech waveform coders reported in the literature use linear predictive coding (LPC) techniques to exploit correlations in the speech signal. Typical examples of these coders are the (baseband) residual excited linear predictive (RELP) coder (Un and Magil, 1975), the adaptive predictive coder (APC) (Atal, 1982), the multipulse excited (MPE) coder (Atal and Remde, 1982) and the regular pulse excited (RPE) coder (Kroon and Deprettere, 1986). All these LPC-based coders produce very good quality speech at bit-rates of 16 kbit/s and above. But, their performance deteriorates significantly for bit-rates lower than 12 kbit/s. This is due to the substantial amount of bits needed for encoding the excitation signal (or the LPC residual).

Atal and Schroeder (1984, 1985) have proposed a Codebook Excited Linear Predictive (CELP) coder for transmitting speech at 4.8 kbit/s. The excitation signal is here encoded using 0.25 bit/sample. The CELP-coder is based on an LPC model of speech and uses the codebook coding (vector quantisation) procedure to encode the excitation signal via an analysis-by-synthesis method. The codebook consists of a set of stochastic codevectors. Encoding of the excitation signal is performed by exhaustively searching the codebook and finding a codevector which minimises an error criterion based on properties of human auditory perception (Schroeder and Atal, 1982; Singhal and Atal, 1984).

Several variants of the CELP coders have been proposed, as the Self Excited Vocoder (Rose and Barnwell, 1986), and the Multiple-Stage Vector Excitation Coder (Davidson and Gersho, 1988).

The aim of this paper is to describe experiments on CELP-based coders, beginning with a refinement of the basic CELP scheme, resulting in several coder variants, and finally evaluating the impact of a realistic channel and background noise on the most interesting coder.

The next section describes the coder algorithms, and presents performance results in segmental signal-to-noise ratios and subjective evaluation by paired comparison (Torgerson, 1958). It is concluded with a ranking of four basic coder models used as a guidance for the issue of quantisation.

Then follows a section concentrating on the development of fully quantised speech coders. The target bit-rate for the coders is motivated by two major international projects attracting large attention. The largest of these aims at developing the half-rate GSM, the Pan European Mobile Telephony System. The other project is for INMARSAT's narrow band service, the INMARSAT-M system. In the light of these systems we have chosen to optimise the coders at two bit-rates, 7.0 kbit/s and 5.0 kbit/s. These bit-rates represent the upper and lower bounds for the two mentioned systems, and also a marked difference in subjective quality. Designs of quantisers for all parameters are described. Finally, a set of 14 coder variants are evaluated, together with 10 reference coders, in a formal subjective listening test, presenting the results in terms of Mean Opinion Scores (MOS) (Kitawaki et al., 1984) for all coders.

In the last two sections the best coder, chosen with regard to subjective quality and complexity, is evaluated under realistic conditions. From experiments with transmission of coded signals over realistic channels, the behaviour of the various parameters in the coder have been classified in accordance to the bit error rate (BER) they can tolerate. Furthermore, the spectral sensitivity to bit errors is investigated, and finally the coder is evaluated in the presence of acoustic background noise typical of mobile environments.

2. Coder algorithms

The Stochastically Excited Linear Predictive coder was first introduced by Atal and Schroeder (1984) as a means of good quality speech coding at very low bit-rates. This chapter describes an implementation of the stochastic coder and various alternative strategies to increase its performance. With reference to Figure 1, the basic CELP algorithm is briefly outlined below.

By a frame-wise linear predictive (LPC) analysis of the incoming speech, the spectral shape of the

signal is modelled using an all-pole filter $1/A(z)$ of order m. At given intervals, the filter coefficients $\{a_k\}$ are estimated from the original speech signal by minimising the energy of the prediction error given by

$$e(n) = x(n) - \sum_{k=1}^{m} a_k \cdot x(n-k). \tag{1}$$

At sub-frame intervals, a long-term correlation analysis (commonly, though imprecisely, known as pitch analysis) is then made on the LPC inverse-filtered signal, by the synthesis filter $1/P(z)$ given in

$$\frac{1}{P(z)} = \frac{1}{1 - \varrho z^{-M}}, \tag{2}$$

where ϱ is the pitch coefficient and M the pitch lag.

Finally, also at sub-frame intervals, an exhaustive search is done through a codebook of random excitation vectors. For each vector C_i a stochastic gain α_i is computed to minimise the energy of the weighted difference signal $E_i(z)$, also expressed as

$$E_i(z) = \frac{A(z)}{A(rz)}\left[S(z) - \alpha_i \frac{C_i(z)}{A(z)P(z)}\right], \tag{3}$$

where $S(z)$ is the original speech. The filter $A(z)/A(rz)$ mimics the auditory masking properties of the human ear, and its use will result in less perceivable signal distortion. A typical value for r is 0.9 (Atal and Schroeder, 1979). The codevector which minimises $\|E_i\|^2$ is then selected, and the parametric representation of the speech signal is thus given by the LPC, the long-term prediction and the codebook coefficients, for subsequent transmission.

Self excitation

The long-term correlation can also be computed using an analysis-by-synthesis procedure (Singhal and Atal, 1984), in which case the coder performance is increased at the cost of added complexity. The long-term correlation lag is restricted to be at or above the sub-block length (typically 40 samples or 5 ms for an 8 kHz sampling frequency), and the correlation is calculated on original and synthesised speech rather than the short-term residual. This is done by searching an adaptive codebook consisting of the most recent excitation vectors, similar to the stochastic codebook search. We have chosen the name of self-excitation for this process, shown in Figure 2. An optimum lag L, is found, and for this the optimum self excitation gain factor, β, is calculated.

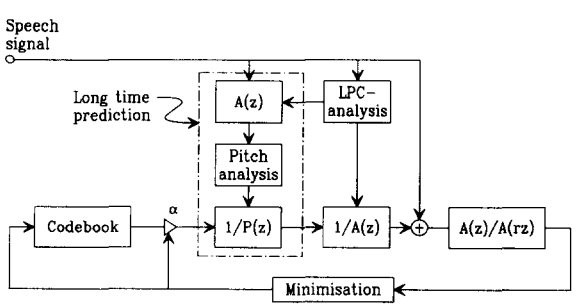

Fig. 1. The stochastic coder (encoder only).

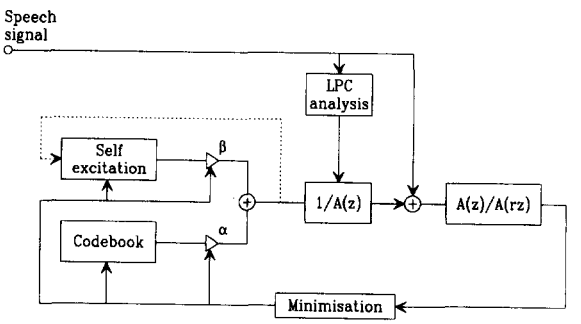

Fig. 2. The self excited stochastic coder (encoder only).

Overlapping vectors

The searches for the optimum codevector and self excitation vector will look similar if the exhaustively-searched stochastic codebook is replaced by a long codebook, i.e., a codebook consisting of overlapping codevectors (Lin, 1986). By this we will gain the combined advantages of both memory savings and a reduction in coder complexity, and it is shown in (Kleijn et al., 1988) that a vector overlap of all but 2 samples for the stochastic codevectors will not affect the speech quality.

Joint optimisation

The optimum codevector and gain factor are traditionally found by an analysis-by-synthesis procedure following the long-term correlation analysis. The performance can, however, be increased by performing a joint optimisation of the self excitation and stochastic gain factors during the codebook search. For each candidate codebook index i, and with a given lag L, the optimum combination of the two gain factors α and β can be found that minimises the energy of the weighted difference between the original and synthetic speech. This can be done by solving a set of linear equations, given e.g. in (Davidson and Gersho, 1988; Kabal et al., 1988). By this, the coder complexity is nearly doubled. The increase in speech quality does, however, make the method interesting, and coders utilising the joint optimisation have been studied.

Stochastic multipulse (or sparse vector) excitation

Motivated by the low bit-rate of the CELP coder and the excellent speech synthesis performed by the higher bit-rate Multipulse (MP) coders (Atal and Remde, 1982), a novel approach combining those to the Self Excited Stochastic Multipulse (SESTMP) coder was suggested in 1987 (Paliwal, 1987), and is described here. The name "Stochastic Multipulse" has been chosen for what is more commonly known as "sparse vectors".

By constructing stochastic codevectors consisting of only a small number of non-zero samples (Davidson and Gersho, 1986) we gain both a reduction in coder complexity and an increase in speech quality. The codevectors are generated by distributing a fixed number of Gaussian, zero-mean samples uniformly over the vector. As opposed to the sub-optimum sequential analysis-by-synthesis procedure in the multipulse coder the pulse positions and amplitudes will be simultaneously optimised under the constraints of a fixed number of pulses in each codevector.

The number of pulses to be contained within a codevector has been subject to investigation; our results are shown in Figure 3, for a 1024-vector codebook, and correspond to those of (Davidson and Gersho, 1986).

It can be seen that the objective speech quality measure (Segmental SNR) increases with the number of pulses up to a certain point where the quality saturates. Informal listening tests have confirmed these results, giving a suitable number of 4 non-zero pulses in each 40-sample codevector.

The coder variants

Several coder variants are to be examined in this paper, and the following naming convention will be used throughout the text. We have chosen the term Stochastic Multipulse for the sparse codevectors, and Self Excitation for adaptive codebooks:

ST:	basic stochastic (CELP) coder;
STMP:	basic coder with stochastic multipulse codebook;
SEST:	self excited stochastic coder; overlapping vectors;
SESTMP:	self excited stochastic multipulse coder; overlapping vectors.

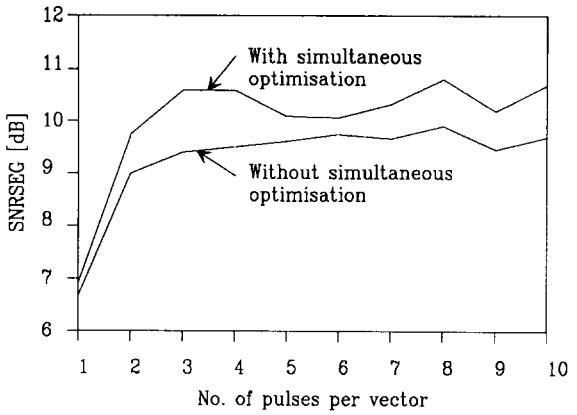

Fig. 3. Objective quality versus number of pulses per vector.

To distinguish between otherwise equal coders, versions using the joint optimisation procedure are marked with an asterisk (*) attached to the coder name.

A comparison of the coder algorithms

To evaluate the performance of the different coder algorithms with respect to each other, the segmental signal-to-noise ratio (SNRSEG) has been measured using an utterance from a Norwegian female speaker. The results are given in Figure 4, where the objective quality is plotted for the ST, SEST, STMP and SESTMP coders, without and with simultaneous optimisation of the gain coefficients. We observe that the use of self excitation in the coders gives a significant quality improvement, and that approxi-

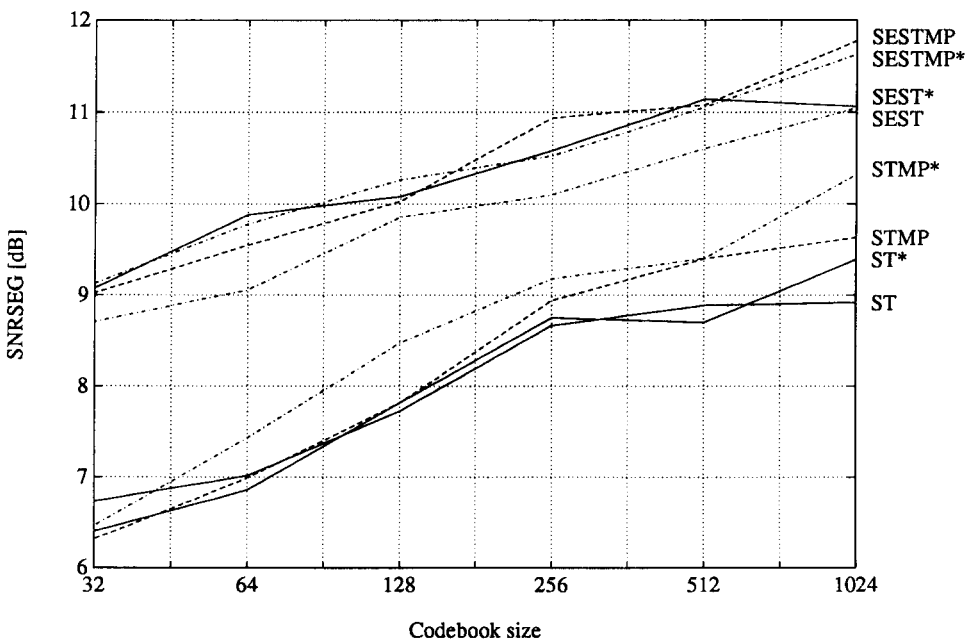

Fig. 4. Coder performance, objective results.

mately 0.5 to 1.0 dB in SNRSEG is gained by the use of Stochastic Multipulse codebooks. The objective quality is increased by a small factor using simultaneous gain optimisation. Informal listening has confirmed these results to a large extent, but the use of joint gain optimisation has a more marked positive effect upon subjective quality.

Postfiltering

The concept of postfiltering is known to improve the subjective quality of various coding schemes, viz. ADPCM (Jayant, 1981; 1987), SBC, ATC, RPE (Perkis et al., 1988). The postfilter is based on both long term and short term spectral information of the speech, representing a re-shaping of the synthesised speech spectrum resulting in a subjective noise suppression. For the stochastic coding schemes the postfilters are generated from the prediction filters at the decoder end, ands thus represent no extra bits to be transmitted.

$$\frac{1}{A'(z)} = \frac{1}{1 - \sum_{k=1}^{m} \eta^k \hat{a}_k z^{-k}}, \quad 0 \leq \eta < 1, \tag{4}$$

$$\frac{1}{P'(z)} = \frac{1}{1 - \varepsilon \hat{\beta} z^{-1}}, \quad 0 \leq \varepsilon < 1, \tag{5}$$

where $\{\hat{a}_k\}$ and $\hat{\beta}$ are the quantised values of $\{a_k\}$ and β. Adding postfiltering with the weighting factors η and ε set to about 0.3 (Kroon and Atal, 1987) will, especially for small codebooks, enhance the subjective coder performance at a small increase in coder complexity. To evaluate the performance of postfiltering, this is added to a few of the coding schemes evaluated in the listening test.

3. Quantisation algorithms

The transmitted information in our parametric coders consists of the LPC-parameters $\{a_k\}$ representing the short term spectral shape, a self excitation sequence or first order pitch predictor taking care of the fine spectral structure, and the codebook index with the gain factor.

The following naming convention will be used in the remaining chapters:

LPC coefficients	$\{a_k\}$
Self excitation gain factor	β
Self excitation lag	L
Pitch coefficient	ϱ
Pitch lag	M
Codebook gain factor	α
Codebook index	i

This section will describe the different parameters, and the design of optimum quantisers for each one of them. This forms a basis for fixing the bit-rates of the fully quantised coders suggested as candidates for the formal subjective listening test presented next. The self excitation lag, pitch lag and codebook index are integer numbers and therefore exactly quantised.

LPC-parameters

The LPC-coefficients $\{a_k\}$, $k = 1, \ldots, m$, calculated using eq. (1), have to be updated, quantised and transmitted for each speech frame.

The speech spectral information in LPC based coders is perceptually important, making quantisation crucial (Kroon and Atal, 1987). Three different quantisation schemes have been applied:

- scalar quantisation of Log Area Ratios (LAR) (Gray and Markel, 1976; Viswanathan and Makhoul, 1985);
- scalar quantisation of Line Spectrum Pairs (LSP) (Soong and Juang, 1984, 1988);
- two stage vector quantisation of LSP (Sundet, 1988).

For the LPC analysis the following parameters are used:

- LPC analysis method: Burg's algorithm;
- Frame length: 20 msec;
- Window length: 22.5 msec;
- Number of coefficients: 10.

Scalar quantisation of reflection coefficients

Efficient quantisation can be obtained by representing the LPC-coefficients by the reflection coefficients. It is generally accepted that a dynamic companding of the coefficients by using the Log Area Ratio-compression characteristic (LAR) can be used with uniform quantisers and give good scalar quantisation of the RC's (Svendsen, 1986).

The LAR parameter l_i is given as

$$l_i = c \cdot \ln \frac{1 - k_i}{1 + k_i}, \tag{6}$$

where k_i is the i-th reflection coefficient (RC) corresponding to LPC-coefficient a_i. The factor c is given by the bias and range defined by the training data. We decided to limit the quantised RC's within the 5–95% area of the histogram, except for RC_1 and RC_2, where the limits were set at 5–99%.

Scalar quantisers for l_i were thus generated, using a speech data base of 102 sentences of Norwegian speech, 17 male and 17 female speakers, recorded in the presence of office background noise. The best (fixed) bit allocation for the quantisation of LAR's is found to be 5,5,5,5,4,4,4,4,2 and 2 bits for coefficients 1 through 10.

The segmental signal-to-noise ratio achieved by this quantisation method is given in Table 2, where it is compared to the line spectrum pair representation (next section).

Line spectrum pairs

The use of line spectrum pairs (LSP) is an alternative representation of the LPC spectral information and is thoroughly described by e.g. Soong and Juang (1984). The LSP frequencies have some important properties such as limited dynamic range, natural ordering and high correlation between coefficients in adjacent frames. In addition, the LSP representation provides efficient control with the shape of the speech spectrum, and allows for a simple check for filter stability. These properties are utilised in designing scalar and vector quantisers for the LSP parameters.

The speech database for optimising the quantisers consist of 244 Norwegian sentences of good quality, 2 male and 2 female speakers, and 102 sentences with office background noise, 17 male and 17 female speakers, giving a total of more than 50000 sets of LPC parameters in the training data base. Three different quantisers have been designed for the LSP frequencies.

Differential scalar quantisation of LSPs

The LPC coefficients $\{a_k\}$, $k = 1, \ldots, m$, are estimated for each frame, and transformed into a set of LSP frequencies, $\{\omega_i\}$, $i = 1, \ldots, m$. Statistical calculations show that there is a high correlation between

coefficients in the same frame as well as between coefficients in adjacent frames. These properties are utilised using simple 1st order prediction schemes from frame to frame as well as within a single frame. The quantiser shown in Figure 5 is based on a closed-loop DPCM scheme with feedback around the quantiser, utilising both inter- and intraframe prediction (Jayant and Noll, 1984). The basic equations describing the quantiser are from Figure 5(a):

$$u_i = \omega_i - \bar{\omega}_i, \tag{7}$$

$$d_i = u_i - h_i \hat{u}_{i-1}, \tag{8}$$

$$d_{ii} = d_i - g_i \hat{d}_i(n-1), \tag{9}$$

where $\bar{\omega}_i$ is the estimated long term mean value of ω_i, $\hat{u}_{i-1}$ is the dequantised value of u from the previous frame with h_i as the estimated prediction coefficient. $\hat{d}_i(n-1)$ is the previous dequantised value of $d_i(n)$ with g_i as the estimated prediction coefficient. $\tilde{d}_i$ and $\tilde{u}_i$ denote the previous quantised value of d_i and u_i, respectively, and $\hat{\omega}_i$ is the dequantised value of ω_i. In the figure the operator T denotes a unit time delay, and D denotes a single frame time delay. Q_i is the quantiser for value i.

Max–Lloyd quantisers (Max, 1960; Perkis and Rowe, 1990) for 0, 1, 2, 3, 4 and 5 bits have been designed for each frequency, and will hereafter be referred to as PRED.

Scalar quantisation of LSP differences

In (Soong and Juang, 1988) it has been observed that the distribution range of LSP frequencies varies significantly with speaker characteristics as well as with recording conditions. To reduce this variability it is proposed to code the difference between adjacent LSP frequencies, instead of the actual LSP frequency itself, motivated by a more suitable distribution range.

LSP difference no. i can be expressed as

$$\Delta\omega_i = \omega_{i+1} - \omega_i, \quad i = 2, \ldots, m, \tag{10}$$

where m is the LPC order. Since $\omega_0 = 0$, LSP frequency no. 1 is coded as it is, while all other LSPs are coded as this difference, henceforth denoted DIF. Optimum Max–Lloyd quantisers (0, 1, 2, 3, 4 and 5 bits) have been designed for each difference.

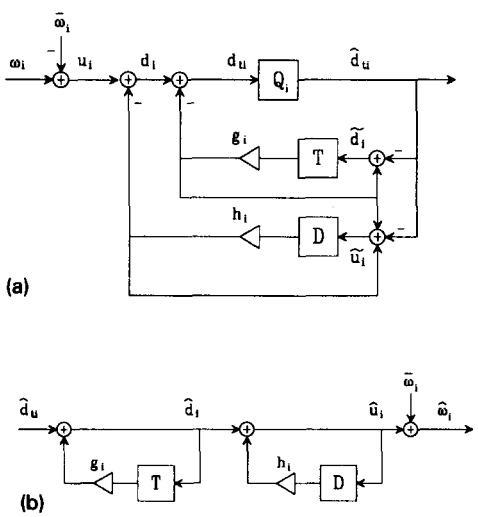

Fig. 5. LSP scalar quantiser with prediction in both time and frequency domain: (a) quantiser; (b) dequantiser.

Vector quantisation of LSPs

In order to remove the redundancy between the coefficients in the same frame even more efficiently, a two stage vector quantisation (VQ) scheme combined with interframe prediction is used (Figure 6). As for PRED the long term estimated mean $\bar{\boldsymbol{w}} = [\bar{\omega}_1, \bar{\omega}_2, \ldots, \bar{\omega}_m]$, is subtracted prior to quantisation. The interframe prediction error, $\boldsymbol{e}$, is given as

$$\boldsymbol{e} = \boldsymbol{u} - \boldsymbol{G}\hat{\boldsymbol{u}}_{i-1}, \tag{11}$$

where $\hat{\boldsymbol{u}}_{i-1}$ are the dequantised values of $\boldsymbol{u}$ from the previous frame and $\boldsymbol{G}$ is a matrix containing the estimated predictor values for the vector on the main diagonal. The two codebooks (VQ1 and VQ2) are searched using the mean square error (MSE) distortion measure.

Two codebooks of size 1024 vectors each were designed by the K-Means algorithm (Linde et al., 1980) using the MSE measure.

Evaluation of LSP-quantisers

In order to evaluate the quantiser performance we have used a spectral distortion measure based on the log root-mean-square (RMS) spectral difference between two given spectra, expressed in dB, given as

$$D = 10\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}(\log_{10}|S(\omega)| - \log_{10}|\hat{S}(\omega)|)^2\,\mathrm{d}\omega} \quad [\mathrm{dB}], \tag{12}$$

where

$$|S(\omega)| = 1/|A(\mathrm{e}^{\mathrm{j}\omega})|^2 \tag{13}$$

and $|\hat{S}(\omega)|$ is the spectrum corresponding to the quantised LPC-coefficients.

Numerical approximations of the spectra are found using a DFT at 256 equidistant frequency points $\omega_i = (\pi \cdot i/256)$, $i = 0, \ldots, 255$.

This measure is chosen because of its high relevance to subjective human perception of speech (Soong and Juang, 1988). It is generally accepted that a spectral distortion less than 1 dB in average is not noticeable, thus 1 dB is set as the perceptually significant difference limen (DL), and quantisers performing below this level are accepted in terms of subjective quality, as long as the error distribution contains few or no large peaks.

The scalar quantisers are evaluated at 40 bits per frame, making them suitable for use in the 7.0 kbit/s coders. The vector quantiser is evaluated at 20 bits per frame with the intended use in 5.0 kbit/s coders. For the scalar quantisation a uniform bit allocation is used. Table 1 gives the mean log RMS spectral distortion for the three schemes calculated over a test database of 924 LPC-sets, taken from outside the training data.

For quantisation of LSP parameters, it is reported that 32 bits per frame achieves the DL criterion.

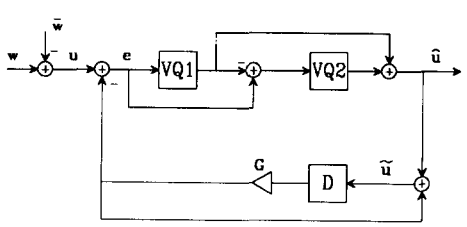

Fig. 6. A two stage cascaded VQ scheme combined with prediction in the time domain.

Table 1
Mean log RMS spectral distortion of three LSP quantisers

Quantiser scheme		D [dB]
SQ of ω_i w/pred	(40 bits/fr)	0.77
SQ of $\Delta\omega_i$	(40 bits/fr)	0.64
VQ of ω_i	(20 bits/fr)	1.39

Both scalar quantisation schemes obviously satisfy the criterion using 40 bits per frame; thus here we have a good opportunity to lower the bit-rate by 8 bits per frame saving 400 bits per second, with little or no decrease in quality. The vector quantisation scheme, however, performs clearly poorer due to allocating only 20 bits per frame, indicating the lower quality of the 5.0 kbit/s coder.

Subjectively the two scalar quantisation schemes differ substantially, the main reason being an entirely different noise characteristic. Quantisation of LSP differences is clearly more "white-noise" like than quantisation of LSPs with prediction, which seems to have a more distinct, rougher kind of noise. Deciding which quantisation scheme to use will therefore be based on personal opinions of speech quality, and may differ from person to person. For a final decision such matters as bit error sensitivity, complexity and behaviour in noisy environments must be considered.

Table 2 gives the segmental SNR for an utterance by a Norwegian female speaker using the SESTMP coder with the different LSP quantisation schemes, and also the scalar quantisation of log area ratios.

Gain factors

In the stochastic coders the excitation signal is built up using a codebook vector with a gain factor in addition to a long term predictor either as a one tap pitch predictor or as a self excitation sequence. The codebook vector and long term predictor model the overall shape of the residual signal, while the gain factors α, ϱ and β give the correct amplitude matching. The excitation signal is given by

$$\text{excit}(n) = \alpha \cdot v(n) + b \cdot c(n), \tag{14}$$

where

$$b = \begin{cases} \varrho, & \text{for pitch predictor calculated by the covariance method,} \\ \beta, & \text{self excitation calculated by analysis-by-synthesis,} \end{cases}$$

$c(n)$ = codebook vector,

$v(n)$ = pitch filter memory or self excitation sequence.

For quantisation purposes the following parameters are used:

- LPC frame length: 20 msec (160 samples);
- Codebook vector size: 5 msec (40 samples);
- $\beta, L; \varrho, M; \alpha, i$ update rate: 5 msec.

We decided to use Max–Lloyd quantisers for the gain factors, using an MSE measure in the design process. For the design a database containing 33000 values each of α, β and ϱ has been created. The values are logged from 9 male and 9 female speakers using 3 utterances from each, recorded in the presence of office background noise. Based on the experimental probability function obtained from the training data, quantisers for 0, 1, 2, 3, 4 and 5 bits were designed for the different gain factors.

According to informal standards in implementing the pitch analysis procedure (the covariance method), the pitch coefficient, ϱ, is limited to the region [0, 1.8]. For a one tap pitch predictor this will provide filter stability, and also simplify the quantiser design by limiting the necessary dynamic range.

Table 2
Segmental SNR using different LPC quantisers (SESTMP)

Quantiser scheme		SNRSEG [dB]
Unquantised		11.78
SQ of LAR		10.51
SQ of ω_i w/pred	(7.0 kbit/s)	11.60
SQ of $\Delta\omega_i$	(7.0 kbit/s)	11.21
VQ of ω_i	(5.0 kbit/s)	8.24

The self excitation gain factor β has an extreme dynamic range due to its peaky nature. Peaks in the range 10–200 times the mean value appear in the transition regions from unvoiced to voiced frames. Due to quantiser design, the dynamic range of β must be limited. In (Paliwal, 1987) this limitation is connected to the self excitation's inherent similarity with the pitch prediction, thereby limiting the dynamic range to [0, 1.8] as for the pitch coefficient. The signal degradation introduced by this limitation is notable and also located in a very sensitive region. Subjectively the limiting (i.e., the quantisation error) in these regions will result in a "blurring like" noise whenever going from silence/unvoiced sounds to speech/voiced regions.

Fully quantised coders

The coder algorithms and quantisation schemes described so far give rise to a large selection of possible coder variants, and for our evaluation four coding schemes have been chosen. These are denoted ST, STMP*, SESTMP and SESTMP*, as is explained in the first section. The four coders operate with equal frame rates as explained earlier, with LPC-frames of 20 ms and excitation subframes of 5 ms. The ST coder uses the autocorrelation method for calculation of the LPC coefficients, while the other coders (STMP*, SESTMP and SESTMP*) estimate the LPC-coefficients by the more complex Burg's algorithm.

Using the available quantisers, two bit-rates were fixed: scalar quantisation of LPC-coefficients giving a bit-rate of 7.0 kbit/s, and vector quantisation suitable for a bit-rate of 5.0 kbit/s. A total description of the parameter choices are given in Table 3.

To be able to make a subjective evaluation of the coders, a formal subjective listening test has been conducted (British Telecom: "Subjective Testing Methodology" (Tro and Stensby, 1988)). Due to practical constraints the maximum number of candidates, including reference coders, was limited to 24.

In our test there were 10 reference conditions as follows:

1 Original.
2–7 Modulated Noise Reference Unit (MNRU) coded utterances with 32, 27, 23, 18, 12 and 6 dB SNR, respectively (CCITT recommendation P70, 1984).
8 The RPE-LTP coder chosen for the GSM mobile telephone system at 13 kbit/s (GSM recommendation 06.10, 1988).
9–10 The APC coder standard for the INMARSAT-B system at 16.0 and 9.6 kbit/s, respectively (INMARSAT Standard-B System Definition Manual, 1988).

This left 14 conditions to be chosen. A short summary of the results used for the selection of coder variants is given below; this has been based on both informal subjective listening and objective results in form of SNRSEG. A thorough description is found in (Ribbum and Perkis, 1988). Most of the selected coders used the bit allocations of Table 3; the exceptions are described is the next section. It is found

Table 3
Parameter choices for the two bit rates (7.0/5.0 kbit/s)

Parameter	7.0 kbit/s coders	5.0 kbit/s coders
LPC-coefficients	40 bits	20 bits
Stochastic gain, α	5 bits	5 bits
Codebook index, i	10 bits	5 bits
Either		
Pitch coefficient, ϱ	3 bits	3 bits
Pitch lag, M	7 bits	7 bits
Or		
Self exc. gain, β	3 bits	3 bits
Self exc. lag, L	7 bits	7 bits

that scalar quantisation of LSPs will give better results than scalar quantisation of LARs. These results are comparable to those reported in (Sundet, 1988). It was, however, found interesting to include at least one coder variant which uses the LAR representation.

Comparing the four coders, it was clear that SESTMP* is the objectively best. Informal listening showed that SESTMP* and SESTMP are of near equal quality and a bit better than STMP*, and also that the stochastic multipulse coders are significantly better than ST. Among the three best coders one must take into account that SESTMP* and STMP* are almost twice as complex as SESTMP.

Regarding 5.0 kbit/s coders, SESTMP and SESTMP* perform equally well and a bit better than STMP*. In the 5.0 kbit/s case ST has not been considered due to its poor quality.

As a last parameter postfiltering (Jayant, 1981, 1987) was considered for the four coding schemes. Postfiltering reduces the general perceived noise level although imposing a clear lowpass characteristic on the synthesised speech. The postfiltering seems to be most advantageous for the low bit-rate coders (5.0 kbit/s).

Although not a standard, a Multipulse coder (MP) (Atal and Remde, 1982) has been included for comparison purposes.

Subjective test

Based on the quality and complexity of the coders, the following 14 conditions were elected as candidates for the formal test:

1	ST	SQ of LSPs	7.0 kbit/s
2	STMP*	SQ of LSPs	7.0 kbit/s
3	SESTMP	SQ of LSPs	7.0 kbit/s
4	SESTMP*	SQ of LSPs	7.0 kbit/s
5	SESTMP*	SQ of LSPs, postfiltered	7.0 kbit/s
6	SESTMP*	SQ of LARs	7.0 kbit/s
7	SESTMP*	SQ of LSPs, 32 entry codebook	6.9 kbit/s
8	SESTMP*	SQ of LSPs, 32 entry, postfiltered	6.9 kbit/s
9	SESTMP*	SQ of LSPs, 128 entry codebook	6.8 kbit/s
10	MP	SQ of LSPs	7.3 kbit/s
11	STMP*	VQ of LSPs	5.0 kbit/s
12	SESTMP	VQ of LSPs	5.0 kbit/s
13	SESTMP	VQ of LSPs, postfiltered	5.0 kbit/s
14	SESTMP	SQ of LSPs, 32 entry codebook	5.0 kbit/s

For coders 7, 8 and 9 the bits made available from the reduced codebook are distributed among the LSP and gain parameters. Coder 14 uses 28 bits for the scalar quantisation of the LSPs, and also fewer bits for the gain factors. The codebooks used for these coders are enhanced versions, resulting from a "pseudo-optimisation" procedure (Ribbum et al., 1988).

Speech material and test method

The speech material consists of hand picked Norwegian utterances read by 2 male and 2 female actors, of typical length 4.0 seconds. The recordings are made using high quality microphones placed in a typical telephone position. The utterances are digitised using 16 kHz sampling with a 7 kHz anti-aliasing filter, and thereafter decimated to 8 kHz sampling using a filter with a cutoff frequency of 3.4 kHz.

Four utterances for each test condition give a total of 96 utterances. In addition, 12 utterances are used for training the test persons. The utterances are presented via standard telephone hand-sets at the

same listening level for 4 test persons simultaneously. Each utterance is judged according to five categories: Excellent, Good, Fair, Poor or Bad.

The test has been performed by 16 persons, 4 groups of 4 persons each. None of the test persons had any prior knowledge to the speech coders, and they were of the age 19–57 years, mainly students 20–25 years of age. All persons have undergone a hearing test, ensuring that they have a listening level over 15 dB in the frequency range 150 Hz = 8000 Hz.

Results

Results from the test are given in Figures 7 to 9. Figure 7 shows the Mean Opinion Score (MOS) (CCITT recommendation P70, 1984) for all 24 conditions used in the test; the MOS for the 6 Modulated Noise Reference Unit (MNRU) values is given in Figure 8, and in Figure 9 the 14 coder candidates (not including the reference coders) are ranged according to their equivalent MNRU values. The latter values are calculated by a linear interpolation between the data points in Figure 8. The suffixes used in the legends should be read as follows:

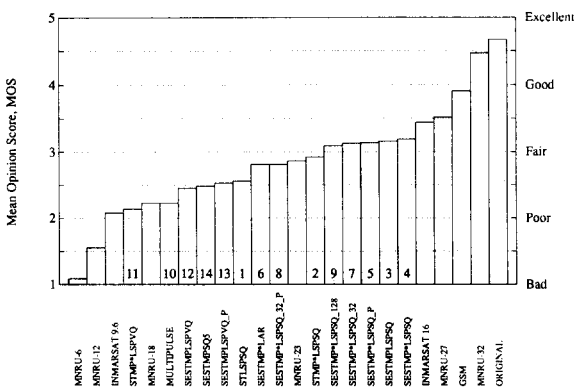

Fig. 7. Mean Opinion Score (MOS) for 24 conditions.

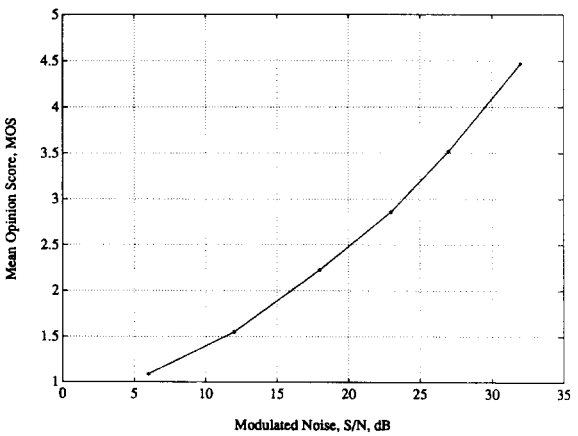

Fig. 8. Mean Opinion Score (MOS) for 6 MNRU values.

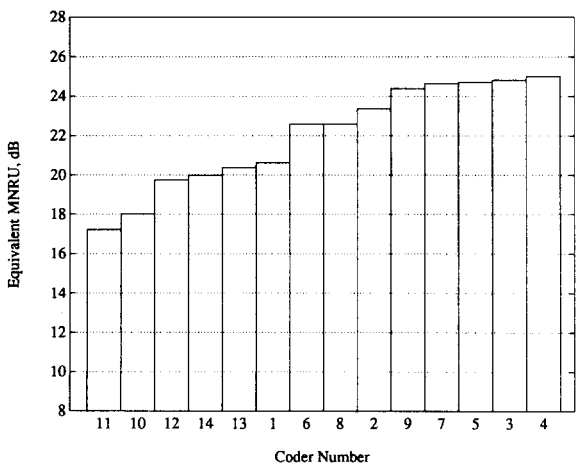

Fig. 9. Equivalent MNRU values for the 14 coder candidates.

LAR – SQ of LPC by LAR (40 bits/frame);
LSPSQ – SQ of ω_i w/pred. (40 bits/frame);
LSPVQ – VQ of ω_i w/pred. (20 bits/frame);
_*nn* – codebook size of *nn*;
_P – postfiltered.

Discussion

The listening test gave several interesting results, making it possible to choose one candidate coder for further work. The earlier assumptions based on several informal tests and objective measures hold. There is an increase in quality as the coders move from the basic ST to a self excited stochastic multipulse coder. Among the latter group of coders, SESTMP(*) at 7.0 kbit/s, there seems to be little or no variation in the MOS (3.09–3.19), while there is a marked reduction in MOS for the vector quantised coders (5.0 kbit/s). Considering these major results, and remembering that SESTMP has half the complexity of SESTMP* (the difference being the simultaneous optimisation of the gains), this coder is chosen for further studies.

As expected, postfiltering gives an improvement in quality for the 5.0 kbit/s coders, while the artifact introduced by the scheme gives a deterioration in quality for the 7.0 kbit/s coders.

Another point worth noting is that all our coders, quality-wise, are placed between the 9.6 kbit/s and 16 kbit/ s INMARSAT standard coders, and also that the coder for the GSM system (13 kbit/s) outperforms all the others. A margin of about 0.2 MOS points should be allowed in the results.

4. Impact of a satellite channel on the 7.0 kbit/s SESTMP

As digital communication has become an increasingly important field, a great deal of research is concentrated on developing new applications. Of these, low bit-rate speech coding for mobile communication using satellites is of special interest. As shown in the previous sections, parametric coders have shown considerable promise for giving high quality speech at bit-rates as low as 5.0 kbit/s. All our coder candidates, however, have thus far been studied in the absence of a realistic channel.

For mobile satellite communication it is necessary to operate within narrow bands at low carrier to noise ratios, in order to keep costs down and make efficient use of the available spectrum. A possible speech coder candidate to such systems could be the Self Excited Stochastic Multipulse (SESTMP) coder with a 7.0 kbit/s bit-rate.

In this section the behaviour of the SESTMP coder over a noisy satellite channel has been studied. The coder is investigated for five characteristic bit error rates, in order to establish the bit error sensitivity of the various parameters. Further, the spectral sensitivity of the two scalar quantisation schemes of the LSPs are discussed. Finally, a set of figures stating the tolerable bit error rates in the SESTMP codec is presented (Perkis et al., 1989; Perkis and Ribbum, 1991).

Description of the speech coder parameters

In the SESTMP coder the input speech is parameterised using 5 basic parameters shown in Table 3, with the pitch coefficient and order not used. This gives the internal structure of one transmitted frame as shown in Figure 10.

Classification of the parameter groups

For the experiment five bit error rates (BER) were chosen: 10^{-4}, 10^{-3}, $4 \cdot 10^{-3}$, 10^{-2} and 10^{-1}. The goal of the test was to establish the sensitivity of the various parameters to random bit errors, both by objective measures (SNRSEG) and subjective preference (paired comparison tests). The speech material used for this testing consists of four Norwegian utterances (2 male, 2 female speakers), each 4.6 seconds long and including natural silence at the end.

Objective measures

Simulations were run separately for each parameter, inserting bit errors according to the appropriate rate, leaving the remaining parameters unchanged. In this way the bit error sensitivity is found for each of the five groups of parameters. Finally, the simulation was run inserting errors in all the parameters (denoted "ALL" in the figures) to obtain the overall performance of the coder.

The segmental SNR (SNRSEG) is calculated for each condition and plotted in a semilog diagram. Figure 11 gives the mean values for the four utterances.

Even though SNR is not a good measure for quality, Figure 11 shows that each parameter alone is hardly affected by BER less than 10^{-3} while the total coder will suffer greatly. The results also show that there is a difference in sensitivity among the codec parameters (Perkis, 1990). It seems clear that the speech spectral information, mainly reflected by the LPC coefficients, is most important. A BER between $7 \cdot 10^{-3}$ and 10^{-2} in the LPC coefficients will result in a total breakdown of the coder, which is heard as large noise bursts and a severe reduction in intelligibility. The self excitation gain factor β is

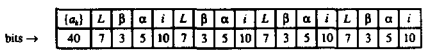

Fig. 10. Internal structure of a 20 msec transmitted frame.

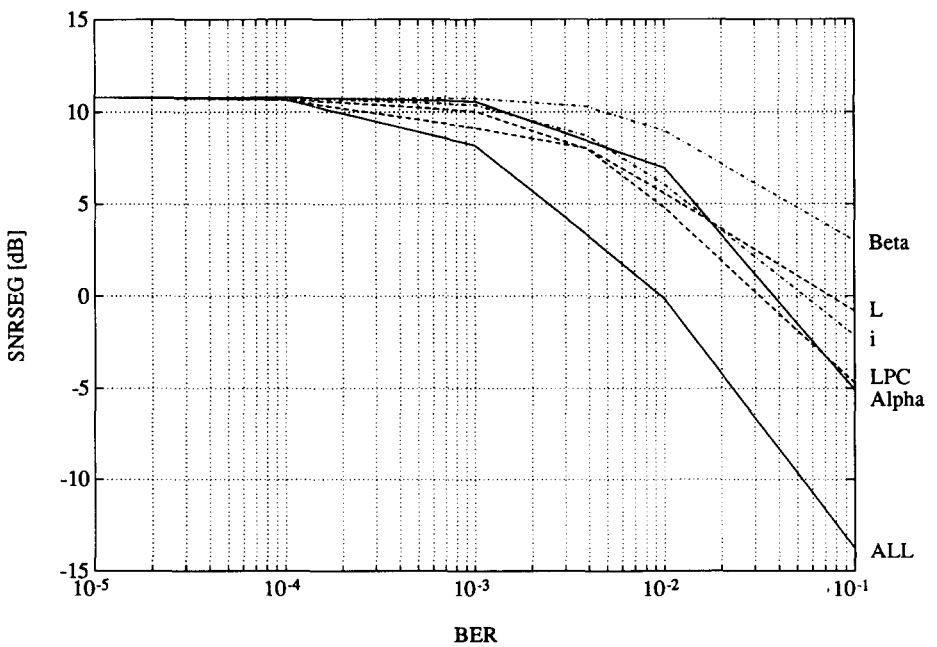

Fig. 11. Objective speech quality (SNRSEG) as a function of bit error rate averaged over four utterances.

the most robust of the parameters and will hardly suffer at all from BER of $5 \cdot 10^{-2}$ or better.

Based on the objective results we obtain the following ranking list of the parameters, from most to least sensitive to bit errors:

$\{a_k\}$, L, α, i, β.

Subjective preference

To evaluate the subjective preference, the severity of the degradations is expected to be difficult to judge because of the vast difference in character as the BER increases. Of the two common methods of evaluating subjective quality the MOS procedure is not preferred here because (i) the method is tedious, and (ii) the quality of the speech differs too much to fit the judgement scale of "excellent"–"bad" in an appropriate way. The paired Comparison Test (PCT), on the other hand, is ideally suited for a small experiment in that it (i) gives the listener a relative measurement scale as opposed to an absolute, and (ii) it is quickly carried out (Sreenivas, 1988).

In the test procedure, among the set of candidate utterances being compared, an exhaustive list of all possible pairs in the set is obtained. This exhaustive list of paired utterances does not have any repetitions, and hence each utterance in the set will be compared only once with every other utterance in the set; also, there are no self comparisons. For n utterances in the set, the paired list has ${}^nC_2$ entries. The order of this list is thoroughly randomised, satisfying two criteria: (i) a particular utterance occurs with equal probability within the list, in both the positions of the pair, (ii) no candidate utterance should appear in the same position of two successive pairs.

Each pair is presented to the listener only once (over headphones), and he/she has to make a "forced-choise" decision. For different listeners, different randomised orders of the list are presented and hence any bias due to a specific random order would get averaged out across listeners. This procedure is a derivation of the recommendation on paired comparison testing adopted by the IEEE sub-committee on subjective measurements (IEEE, 1969).

For most purposes the mean preferential score for each utterance averaged over all listeners is a sufficient indication. However, we have chosen to present the subjective preference as a ranking on a psychological scale, based on a simple Gaussian statistical model (Torgerson, 1958), where 0.0 denotes the highest possible score.

To be able to further validate the ranking of parameter sensitivity obtained by the objective results, we have run 4 different PCTs. For our purpose we have chosen to subject the first female utterance to BERs of 10^{-4}, 10^{-3}, $4 \cdot 10^{-3}$ and 10^{-2}. Using a single utterance only will, especially at the lower BERs, give results of somewhat poor statistical significance. However, it is still believed that the tests will give valid results at a BER of 10^{-2}, which will be used for comparison with the objective results. The rest of the BER are presented for completeness, and to be able to discuss the dependence on error location.

Figure 12 gives the relative sensitivity among the parameters. The results clearly verify the objective results, in that β is the least sensitive parameter. The most sensitive parameters are the LPC coefficients and α. Due to the distinctly different error tolerance among the parameters, the error location will be of major subjective significance, especially at low BER. This can be seen in Figure 12 as a marked improvement in performance as the BER goes from 10^{-3} to $4 \cdot 10^{-3}$. Specifically, the counter-intuitive trend seen in the LPC parameters (and also for L) at this point can be considered a result of this. At this BER a total of 37 errors are inserted in a test data set of 9200 bits. Considering the low bit error rate, we assume that every parameter of 4 bits will only be affected by one error at a time. The quality degradation caused by this error will differ enormously according to the positioning of this error, as shown in the figure. At higher error rates, however, the errors will be more evenly distributed, and the quality degradation more predictable.

One way of avoiding these discrepancies in the results would be to average the performance over several runs with different error distributions at the same rate. This would decrease the rather large

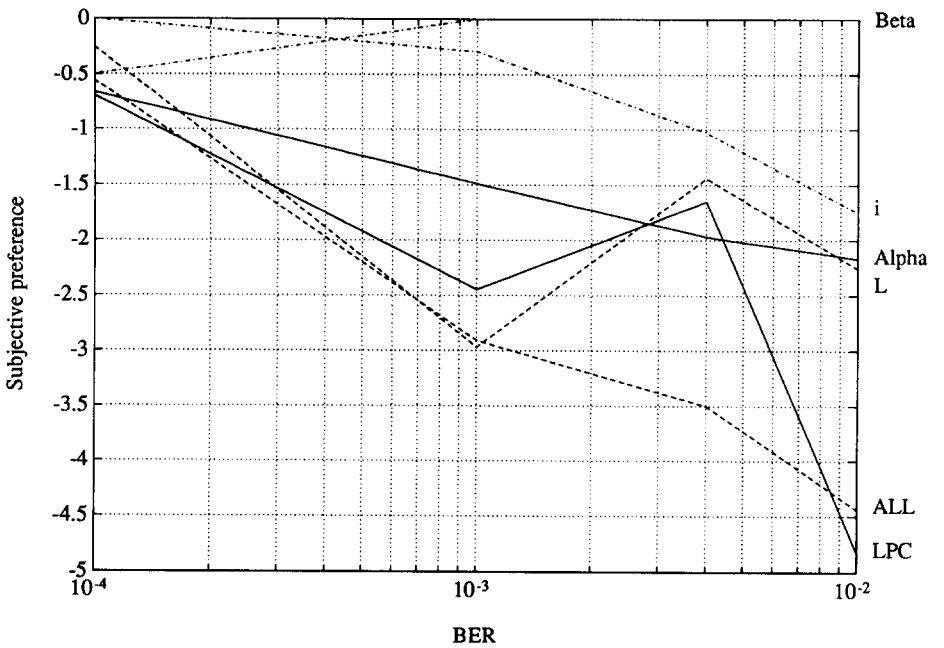

Fig. 12. Subjective rating of the relative importance of bit errors in the various parameters.

statistical errors, indicated in our tests by a high standard deviation, and give a stronger reflection on the differences in performance of the parameters at the low BERs.

However, our purpose of the test has been successfully obtained. The results show the same ranking as the objective results, which clearly verifies that the critical point is somewhere between 10^{-3} and 10^{-2}. Worth noting is the subjective effect of the bit errors. By inserting errors in β, L and i the speech seems more noisy and feels harder to understand in a sense of harshness, while errors in the LPC coefficients and α result in serious degradations of the spectrum and dynamic range. The errors result in gross distortions and large amplitudes, some of which are painful to the ear. Some of these errors are due to the poor correction algorithm for unstable filters.

Correcting unstable filters

Representing the LPC-coefficients by LSP frequencies has certain advantageous properties with regard to quantisation. One of those is the possibility of conserving the filter stability after quantisation, by observing that the LSP frequencies (ω_1–ω_{10}) are interlaced, i.e., $\omega_1 < \omega_2 < \cdots < \omega_{10}$.

During the quantisation process, closely spaced LSP frequencies may change order, resulting in an unstable synthesis filter (this may occur with the PRED method only). A similar problem will occur if the dequantisation moves the LSP frequencies close together, resulting in a large resonance in the synthesised speech spectrum. A simple algorithm for avoiding these problems is suggested in (Sundet, 1988). This algorithm is capable of correcting unstable filters only if two consecutive LSP frequencies are interchanged, and not in the case if ω_i and ω_{i+j}, $j > 1$, would interchange, as can happen in the presence of bit errors, i.e., it will not guarantee the ordering stepping through the algorithm once.

A new algorithm for preserving filter stability is therefore proposed (Perkis and Ng, 1990), described in Figure 13, which involves a repeated ordering procedure referred to as sorting, where $\underline{\hat{\omega}}_{DQ}$ refers to the dequantised values of $\{\omega_i\}$, and $\underline{\hat{\omega}}$ refers to the set after a possible modification. This ensures that the LSPs are monotonically increasing.

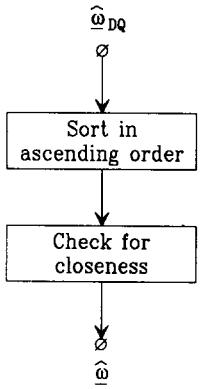

Fig. 13. Algorithm for correction of unstable filters.

Using this algorithm we observe an increase in SNR for BER = 10^{-2} and 10^{-1} for the LSP parameters. Subjectively the improvement is remarkable for these BERs. There is still a large distortion, but as expected the spectral peaks are smaller. The painful distortion was to a large extent due to instability of the synthesis filter, and has now been reduced. At 10^{-1}, however, the coder still breaks down and if not directly painful is definitely unpleasant to listen to.

Spectral error sensitivity of LSP frequencies

In this section we will further investigate spectral sensitivity to bit errors as well as the sensitivity within a set of LPC coefficients. Two different quantisation schemes have been evaluated: scalar quantisation of LSP frequencies with prediction both in time and frequency (PRED) and scalar quantisation of LSP differences (DIF). The spectral sensitivity of the two quantisers are considered both by objective and subjective means. Finally the trade-offs inherent in choosing one of the quantisation schemes are discussed.

To be able to compare the two quantisation schemes with respect to bit errors, both SNRSEG and informal listening have been taken into account. Figure 14(a) shows the SNRSEG calculated for each of the 5 BER and plotted in semilog diagrams. From the figure it seems that DIF is somewhat more robust, and can tolerate BER as high as 10^{-2}. Figure 14(b) gives the subjective results, where 0.0 is the highest score.

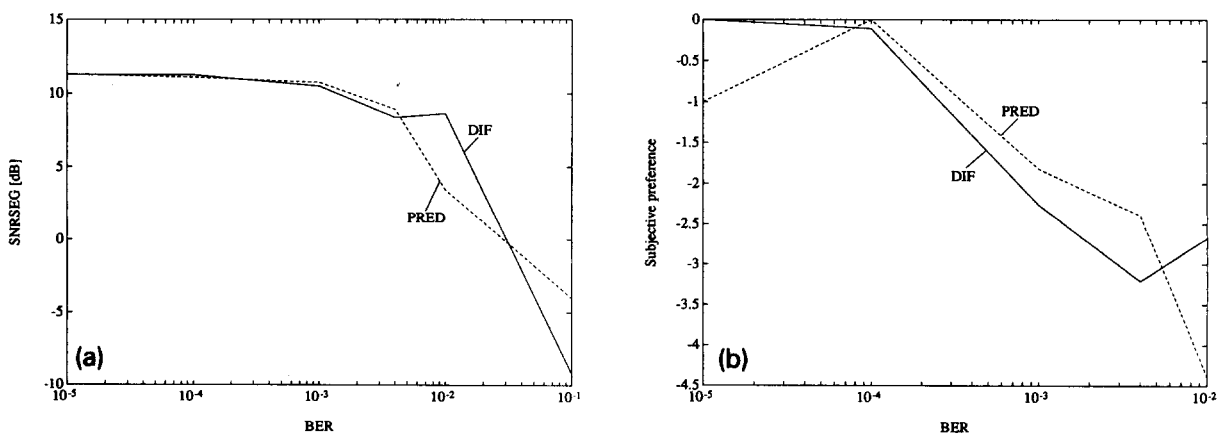

Fig. 14. Comparison of PRED and DIF as a function of BER.

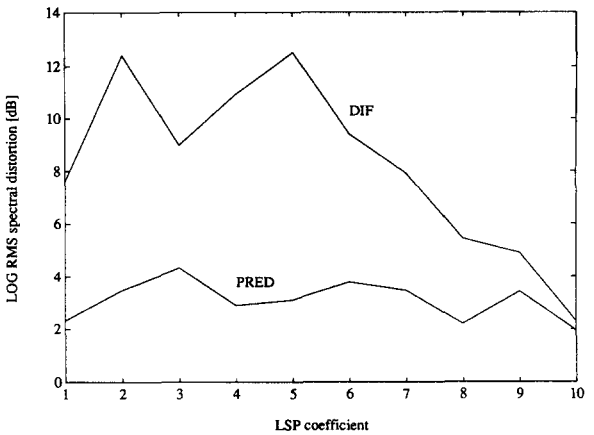

Fig. 15. Log RMS spectral distortion for bit errors in MSB of ω_1–ω_{10}.

In order to derive any sensitivity difference within a set of LSP parameters an error is introduced in the Most Significant Bit (MSB) of each parameter within the set, and the log RMS spectral distortion is calculated. Figure 15 gives the results for the 10 LSP frequencies and differences. The results are averaged over 4 different speech frames from a female utterance, two voiced, one unvoiced and one transition frame, respectively.

Discussion

As stated earlier, errors in PRED can give large peaks in the spectrum. These peaks occur when an error moves the LSP frequency ω_i too close to $\omega_{i\pm1}$ creating a resonance in the spectrum. It is also clear that an error within a frame is isolated to a small region. Inherent in the prediction scheme is that an error in ω_i will influence ω_k, $k = i, \ldots, m$. In addition the predictors for the next frames will suffer. In other words an error in PRED will give degradation in an isolated region within a frame, but will to a great extent be spread out in time.

This again can give large discontinuities in the spectral estimate from frame to frame partly giving rise to the previously mentioned painful distortion.

In quantising the LSP differences, an error will generally spread out in the whole frame. An error in difference $\Delta\omega_i$ will affect all the differences $\Delta\omega_k$, $k = i + 1, \ldots, m$. Thus, an error increasing the LSP difference will shift the spectrum to the right. LSP frequency $m + 1$ is a constant (π), thus a large positive error in $\Delta\omega_i$ will press the spectrum towards π, and could give a resonant at the high frequencies. Analogously an error decreasing the LSP difference will shift the spectrum to the left.

Therefore it seems clear that the LSP differences $\Delta\omega_1$–$\Delta\omega_{10}$ will have a decreasing sensitivity with increasing index. Figure 15 supports this fact for four characteristic frames from a female utterance. In general, Figure 15 shows peaks in log RMS spectral distortion in the area one would expect to find the two first formants in human speech (ω_2–ω_4, ω_5–ω_7) both for PRED and DIF.

In PRED there seems to be no obvious sensitivity difference within the set of LSPs. Subjectively the two quantisation schemes perform differently. Errors in DIF give a higher noise level and a greater loss of intelligibility in the speech, but far less occurrence of large peaks. Figure 14(b) shows this difference at BER = 10^{-2}. While PRED at this BER give painful cracks, DIF still gives tolerable speech. In Figure 14(a) also, DIF shows a significant higher SNR for BER = 10^{-2} than PRED.

To conclude, it is clear that there exist several trade-offs in selecting one of the quantisation schemes:

- Subjectively, without errors, the two quantisation schemes perform similar, giving only different noise characteristics. Personal preferences will differ.

- In the presence of bit errors, PRED gives large peaks in the spectrum and an often painful distortion. Intelligibility, however, is well kept. Errors are isolated to a small region in a single frame, but will influence on several frames due to prediction.
- DIF generally gives a higher overall noise level, and a greater loss of intelligibility. The large peaks, however, occur far less frequently. An error is generally spread out in the whole frame, but will not affect subsequent frames.
- Within a frame, the LSP differences seem to have decreasing spectral sensitivity with increasing i ($i = 1, \ldots, m$).
- In terms of complexity, DIF is the simplest quantisation scheme.

Tolerable BER for the parameter groups

Based on the results in this section, where the objective and subjective speech quality is given as a function of BER, three classes of quality degradation are proposed in order to classify the parameter groups:

(i)	worst case	indicating the maximum BER tolerated on each parameter, indicating the lower limit of acceptable subjective quality;
(ii)	mean	indicating a BER giving the subjective quality tolerable "most of the time",
(iii)	limiting case	indicating the maximum BER giving no perceivable degradation of the speech.

The resulting BERs for these degradations are given in Table 4.

The error sensitivity varies for the different bits of the parameters, and they will need different degrees of protection. Based on further tests, and also on the afore-mentioned results, a classification on bit level has been obtained, though not discussed here.

5. Acoustic background noise

In the use of LPC based speech orders, estimation of the spectrum is a key point. Traditionally the simulations have been carried out using clean speech, i.e., speech data recorded in silent rooms under controlled speech levels. A problem with the mobile environment is the background noise producing noisy data at the coder input. In this section the SESTMP codec is evaluated with regards to its robustness under 5 typical mobile conditions.

Speech data with noise

The sentences used as speech data are excerpts from different conversations from experiment E166 of the Pan European Mobile Radio Conversation Test (Table 5). The sentences are recorded in moving

Table 4
Tolerable BER for the five parameter groups

Parameter	BER worst case	BER mean case	Ber limiting case
$\{a_k\}$	$1 \cdot 10^{-3}$	$6 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
L	$3 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$
β	$2 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$5 \cdot 10^{-3}$
α	$1 \cdot 10^{-3}$	$6 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
i	$3 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-3}$

Table 5
Experimental conditions from the E166 Pan European Mobile Radio Conversation Test, chosen for our studies

Condition	Voice	Description
1	F2	Lorry in urban area, variable speed with mean 50 km/h, open window
5	M1	Small car in rural area, constant speed 100 km/h, closed window
10	M2	Large car in rural area, constant speed 100 km/h, closed window
11	F1	Large car in urban area, variable speed with mean 50 km/h, open window, very low voice level

Table 6
Performance of SESTMP in the presence of acoustic background noise

Condition	SNRSEG [dB]
1	9.21
5	7.42
10	7.67
11	4.36

cars equipped with summer tyres on a concrete road surface. Four different noise conditions are chosen for our study; two female (abbreviated F1, F2) and two male (M1, M2) voices were recorded under the conditions listed.

Quantiser performance

The quantisers PRED and DIF were optimised using a data base consisting of clean speech as well as speech with office background noise. Table 6 shows the objective results (SNRSEG) for the 4 chosen noisy sentences using PRED. The bit-rate of the codec is fixed at 7.0 kbit/s.

Despite the expected quantiser mismatch, the coder performs well under these conditions. The low SNRSEG reflects the higher noise level present in this speech, and does not reflect the actual subjective quality. The extremely low SNRSEG for condition 11 reflects the low input level at the coder. The coded

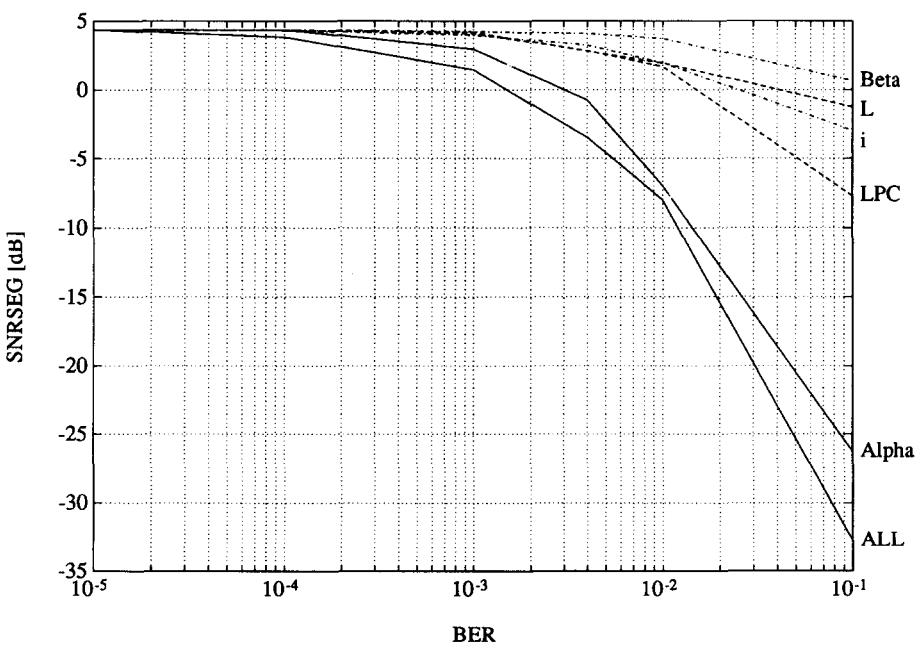

Fig. 16. Segmental SNR as a function of bit error rate for noise condition 11.

speech is still intelligible although the noise level is high. In the other conditions the coder seems to give relatively better quality than for clean speech. This is most likely due to masking effects; the quantisation noise is efficiently masked by the already existing background noise.

Another point worth noting is the coder's ability to code the background noise. The background sounds are easily perceivable; an accelerating car still has the characteristic sound known so well. This is an important aspect in the user comfort of a mobile service.

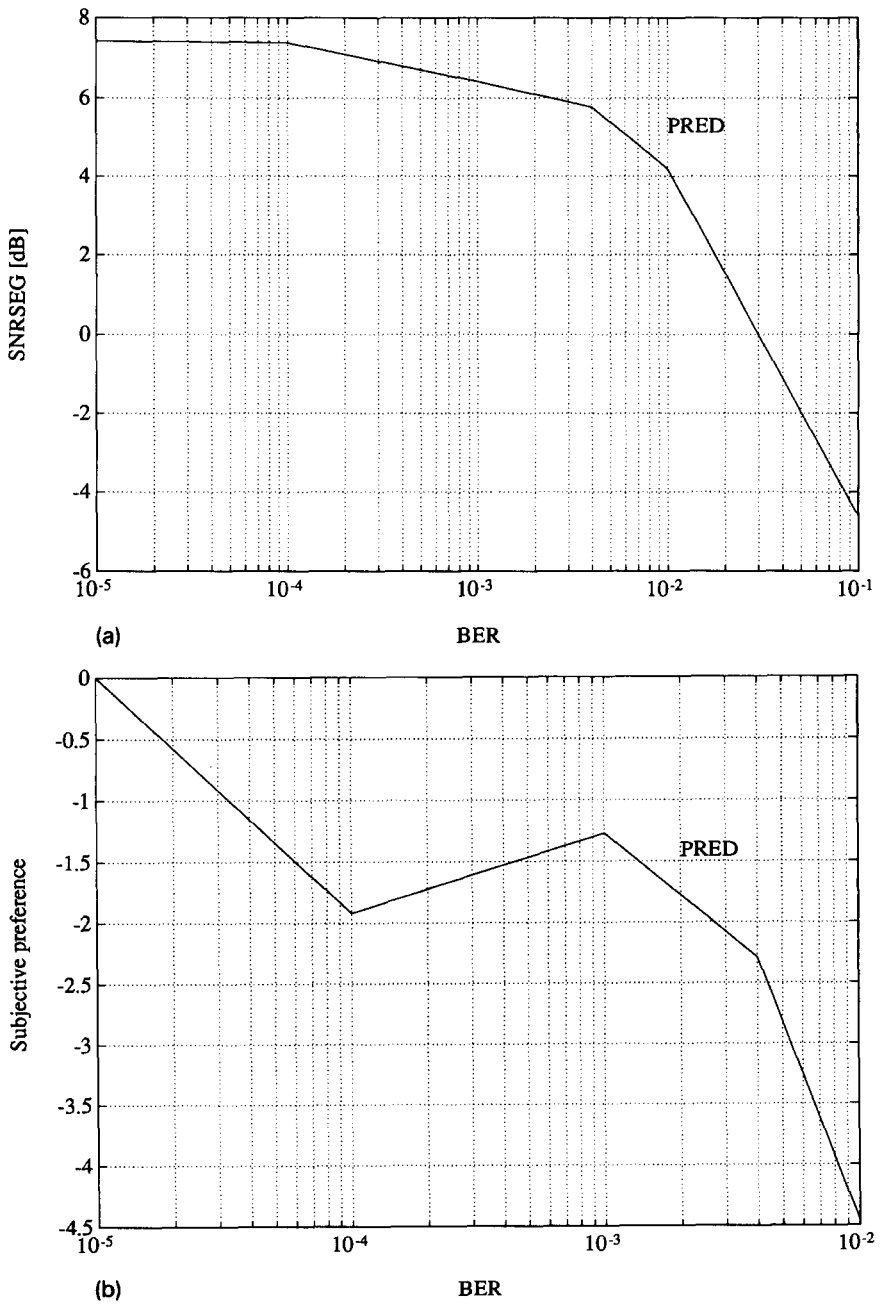

Fig. 17. Bit error sensitivity of LPC-coefficient for noise condition 5.

Bit error sensitivity

The simulations for establishing the bit error sensitivity in the presence of acoustic background noise were run on an estimated worst case sentence, condition 11 above. This is a long female utterance (approximately 37 seconds) with a very low speech level.

The SNR is calculated for each of the five parameter groups and BER conditions and plotted in a semilog diagram shown in Figure 16.

The figure basically shows the same trend as reported for clean speech, only a bit more robust. The background noise masks out much of the noise caused by bit errors. The stochastic gain factor, α, however, is more sensitive, a fact that may be due to the low input level.

To subjectively establish the spectral sensitivity to errors, another sentence, condition 5, was chosen. Figure 17(a, b) shows the effects of bit errors on the LSP frequencies using PRED. These results again are quite similar to the ones reported for clean speech. Again one can clearly note the subjective importance of the error locations, discussed in conjunction with Figure 12, as the quality seemingly increases when the BER goes from 10^{-4} to 10^{-3}.

The conclusion is that SESTMP performs as well (if not better) in the presence of acoustic background noise as on clean speech. This gives a satisfactory robustness and the coder is well suited for use in a mobile communication system.

6. Conclusion

The LPC model is successful in modelling speech waveforms, and serves as a basis for several speech coders at varying bit-rates, ranging from 16 kbit/s downwards. In this article several variations of the Code Excited Linear Predictive (CELP) coder have been discussed, and the difference in speech quality between the versions has been examined. In general, the CELP coders are able to yield good quality speech at low bit-rates, but does so at a relatively high computational cost.

The alterations to the original CELP coder suggested here will both improve the speech quality and simplify a real-time implementation. A noticeable quality enhancement is attained by introducing a self excitation sequence to replace the long term (or pitch) predictor. Both a quality gain and a coder simplification results from the use of a Stochastic Multipulse codebook consisting of only few non-zero samples in the vectors, and the best results are reported from a coder applying both these methods. Further improvements in quality have been observed by performing a joint optimisation of the excitation gain factors; this does, however, imply a more complex coder.

Quantisers have been generated and evaluated for all the different coder parameters, and possible bit allocation schemes are proposed for coders at 7.0 and 5.0 kbit/s. For the LPC parameters, different quantisation schemes are covered and compared.

The fully quantised coders have been evaluated in a formal subjective test, resulting in Mean Opinion Scores (MOS) and equivalent MNRUs.

A Self Excited Stochastic Multipulse coder is suggested as suitable for use in mobile communication systems. This coder at 7.0 kbit/s has been studied against different bit error rates in the parameters, and both objective and subjective criteria are used in an evaluation of bit error sensitivity. Finally, the performance of this coder in the presence of background noise is discussed, and the coder is found to be very robust against vehicle-generated background sound.

Several interesting topics concerning CELP-based speech coders have not been discussed in this article. To aid a real-time implementation, a number of methods exist, like the use of ternary codebooks (Salami and Appleby, 1989) or general simplification methods for codebook search (Trancoso and Atal, 1986). Further quality improvements can be achieved by calculating the long-term predictor at sub-sample intervals (Kroon and Atal, 1989) and several other methods. If the coder delay were not a problem,

like in a store-and-forward system, an analysis done over longer speech intervals could result in significantly better quality/rate performance.

Acknowledgments

This work has been supported by the Multiclient Project in Speech Processing at ELAB-RUNIT, The Norwegian Institute of Technology, Trondheim, Norway.

References

B.S. Atal (1982), "Predictive coding of speech at low bit rates", *IEEE Trans. Comm.*, Vol. COM-30, pp. 600–614.

B.S. Atal and J.R. Remde (1982), "A new model of LPC excitation for producing natural-sounding speech at low bit-rates", *Proc. Internat. Conf. Acoust. Speech Signal Process., Paris, 1982*, pp. 614–617.

B.S. Atal and M.R. Schroeder (1979), "Predictive coding of speech signals and subjective error criteria", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-27, No 3, pp. 247–254.

B.S. Atal and M.R. Schroeder (1984), "Stochastic coding of speech signals at very low bit-rates", *Proc. ICC, Amsterdam, 1984*, pp. 1610–1613.

British Telecom, CSELT and Swedish Telecommunication Administration: "Subjective testing methodology", *CEPT TR3/COST 207*, Doc. no. 86/8 Revised 1.

CCITT recommendation P70, *Modulated Noise Reference Unit*, 1984.

G. Davidson and A. Gersho (1986), "Complexity reduction methods for vector excitation coding", *Proc. Internat. Conf. Acoust. Speech Signal Process., Tokyo, 1986*, pp. 3055–3058.

G. Davidson and A. Gersho (1988), "Multiple-stage vector excitation coding of speech waveforms", *Proc. Internat. Conf. Acoust. Speech Signal Process., New York, 1988*, pp. 163–166.

A.H. Gray and I.D. Markel (1976), "Quantization and bit allocation in speech processing", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, pp. 459–473.

GSM recommendation: 06.10, Version 3.1.2, 10 September 1988.

"IEEE recommended practice for speech quality measurements", *IEEE Trans. Audio Electroacoust.*, September 1969, pp. 227–246.

INMARSAT Standard-B System Definition Manual, Module 1, App. 1, "Voice Coding System Description", 9 May 1988.

N.S. Jayant (1981), "Adaptive post-filtering of ADPCM speech", *Bell Syst. Techn. J.*, Vol. 60, No. 5, pp. 707–717.

N.S. Jayant (1987), "ADPCM coding of speech with backward adaptive algorithms for noise feedback and postfiltering", *Proc. Internat. Conf. Acoust. Speech Signal Process., Dallas, 1987*, pp. 1288–1291.

N.S. Jayant and P. Noll (1984), *Digital Coding of Waveforms: Principles and Applications to Speech and Video* (Prentice-Hall, Englewood Cliffs, NJ).

P. Kabal, J.-L. Moncel and C.C. Chu (1988), "Synthesis filter optimization and coding: Applications to CELP", *Proc. Internat. Conf. Acoust. Speech Signal Process., New York, 1988*, pp. 147–150.

N. Kitawaki, M. Honda and K. Itoh (1984), "Speech quality assessment methods for speech coding systems", *IEEE Comm. Mag.*, Vol. 22, No. 10, pp. 26–33.

W.B. Kleijn, D.J. Krasinsky and R.H. Ketchum (1988), "Improved speech quality and efficient vector quantization in SELP", *Proc. Internat. Conf. Acoust. Speech Signal Process., New York, 1988*, pp. 155–158.

P. Kroon and B.S. Atal (1987), "Quantization procedures for the excitation in CELP coders", *Proc. Internat. Conf. Acoust. Speech Signal Process., Dallas, 1987*, pp. 1649–1652.

P. Kroon and B.S. Atal (1989), "On improving the performance of pitch predictors in speech coding systems", *IEEE Workshop on Speech Coding for Telecommunications, Vancouver, 1989.*

P. Kroon and E.F. Deprettere (1986), "Regular-pulse excitation – A novel approach to effective and efficient multipulse coding of speech", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-34, pp. 1054–1063.

D. Lin (1986), "New approaches to stochastic coding of speech sources at very low bit rates", in *Signal Processing III: Theories and Applications*, ed. by I.T. Young, A.P.W. Duin, J. Biemond and J.J. Gerbrands (North-Holland, Amsterdam).

Y. Linde, A. Buzo and R.M. Gray (1980), "An algorithm for vector quantization design", *IEEE Trans. Comm.*, Vol. COM-28, No. 1, pp. 84–95.

I. Max (1960), "Quantizing for minimum distortion", *IEEE Trans. Information Systems.*

K.K. Paliwal (1987), "Stochastic, multipulse and self excited linear predictive coders for low bit-rate coding of speech", *ELAB Report STF44 F87108.*

A. Perkis (1990), "Speech coding for mobile satellite communications; A novel scenario", *Proc. ISSPA, Brisbane, 1990*, pp. 710–714.

A. Perkis and T.S. Ng (1990), "A robust low complexity 5.0 kbps stochastic coder for a noisy satellite channel", *Proc. TENCON '90, Hong Kong*, pp. 334–338.

A Perkis and B. Ribbum (1991), "Application of stochastic coding schemes in satellite communications", in *Advances in Speech Coding*, ed. by B.S. Atal, V. Cuperman and A. Gersho (Kluwer Academic Publishers, Dordrecht), pp. 277–286.

A. Perkis and D. Rowe (1990), "Quantiser design and evaluation procedures for hybrid voice coders", *Proc. 3rd Australian Internat. Conf. on Speech Science and Technology, Melbourne, 1990*, pp. 40–46.

A. Perkis, B. Ribbum and I.J. Fanneløp (1989), "A study on the impact of a satellite channel on a 7.0 kbit/s CELP based coder", *IEEE Workshop on Speech Coding, Vancouver, 1989.*

A. Perkis, B. Ribbum and T.A. Ramstad (1988), "Improving subjective quality in waveform coders by the use of postfiltering", *Proc. 2nd Australian Internat. Conf. on Speech Science and Technology, Sydney, 1988*, pp. 60–65.

B. Ribbum and A. Perkis (1988), "Fully quantized CELP-based speech coders and codebook enhancements", *ELAB Report STF44 F88152.*

B. Ribbum, A. Perkis and K.K. Paliwal (1988), "Enhancing the codebook for improving the speech quality of CELP coders", *Proc. 2nd Australian Internat. Conf. on Speech Science and Technology, Sydney, 1988*, pp. 408–413.

R.C. Rose and T.P. Barnwell (1986), "The self excited vocoder – An alternate approach to toll quality at 4800 bps", *Proc. Internat. Conf. Acoust. Speech Signal Process., Tokyo, 1986*, pp. 453–456.

R.A. Salami and D.G. Appleby (1989), "A new approach to low bit rate speech coding with low complexity using binary pulse excitation (BPE)", *IEEE Workshop on Speech Coding for Telecommunications, Vancouver, 1989.*

M.R. Schroeder and B.S. Atal (1982), "Speech coding using efficient block codes", *Proc. Internat. Conf. Acoust. Speech Signal Process., Paris, 1982*, pp. 1668–1671.

M.R. Schroeder and B.S. Atal (1985), "Code-excited linear prediction (CELP): High-quality speech at very low bit rates", *Proc. Internat. Conf. Acoust. Speech Signal Process., Tampa, 1985*, pp. 937–940.

S. Singhal and B.S. Atal (1984), "Improving performance of multipulse LPC coders at low bit rates", *Proc. Internat. Conf. Acoust. Speech Signal Process., San Diego, 1984*, pp. 1.3.1–1.3.4.

F.K. Soong and B.-H. Juang (1984), "Line spectrum pairs (LSP) and speech data compression", *Proc. Internat. Conf. Acoust. Speech Signal Process., San Diego, 1984*, pp. 1.10.1–1.10.4.

F.K. Soong and B.-H. Juang (1988), "Optimum quantization of LSP parameters", *Proc. Internat. Conf. Acoust. Speech Signal Process., New York, 1988*, pp. 394–398.

T.V. Sreenivas (1988), "Modelling LPC-residue by components for good quality speech coding", *Proc. Internat. Conf. Acoust. Speech Signal Process., New York, 1988*, pp. 171–174.

A.F. Sundet (1988), "Efficient quantization of speech spectral information using line spectrum pairs (LSP)", *Diploma thesis, The Norwegian Institute of Technology*, Trondheim, Norway.

T. Svendsen (1986), "Modellvalg for APC", *ELAB project memo AN86198* (in Norwegian).

W.S. Torgerson (1958), *Theory and Methods in Scaling* (Wiley, New York).

I.M. Trancoso and B.S. Atal (1986), "Efficient procedures for finding the optimum innovation in stochastic coders", *Proc. Internat. Conf. Acoust. Speech Signal Process., Tokyo, 1986*, pp. 2375–2378.

J. Tro and S. Stensby (1988), "Subjective evaluation of CELP based coders", *ELAB Report STF44 F88175* (in Norwegian).

C.K. Un and D.T. Magil (1975), "The residual excited linear prediction vocoder with transmission rate below 9.6 kbit/s", *IEEE Trans. Comm.*, Vol. COM-23, pp. 1466–1474.

R. Viswanathan and I. Makhoul (1985), "Quantization properties of transmission parameters in linear predictive systems", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, pp. 309–321.