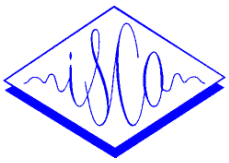

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

HIGH QUALITY 14.1kb/s WIDEBAND SPEECH CODER

A. W. Black, I. A. Atkinson, A. M. Kondoz, B. G. Evans

Centre for Satellite Engineering Research
University of Surrey, Guildford, Surrey, GU2 5XH, U.K.
Tel: +44 1483 259803, Fax +44 1483 259504
e-mail: eep2ab@ee.surrey.ac.uk

ABSTRACT

This paper describes the application of the Pulsed Residual Excited Linear Prediction (PRELP), a variation of the well known CELP algorithm to coding 7kHz wideband speech. The coder is to be used in the audio description of television. The PRELP algorithm operates at the source rate of 14.1 kb/s and uses an innovative excitation which is adapted for wideband speech. In addition to producing high quality speech, the algorithm simplifies the complexity of the encoder/decoder. This ensures that a cost effective hardware implementation can be achieved. The coder also employs a novel codebook gain quantiser, which is adaptive to the energy of the LPC synthesis filter response.

1. INTRODUCTION

The speech coding algorithm is to be used for high quality audio description of TV programmes for the partially sighted. This leads to number of constraints to which the algorithm must adhere. The complexity of the decoder has to be kept as simple as possible in order to produce inexpensive television receiver units. The decoder has to work with 0.1% channel errors without significant loss of quality in the decoded speech. This performance can easily be achieved using an adequate level of Forward Error Correction (FEC) for bit protection. However, the overall bit rate is limited by the spare capacity of 2 lines of the Teletext system at 15.2 kb/s. In addition to the limitation of the bit rate, the frame size of the system is determined by the system clock which is fixed at 20 ms.

2. PRELP SPEECH CODING

The coder used for this application is based on the widely reported CELP [1] algorithm. It is designed to operate with frame and subframe update rates of 50 and 400Hz respectively. A 16th order LPC analysis is performed over a 20 ms frame and coded using Line Spectral Frequencies (LSF). There is an additional half frame lookahead for linear interpolation of the LSFs. Long Term Prediction (LTP) and secondary excitation analysis is performed over the full band of the signal.

The principal difference between the PRELP and CELP algorithms is the construction of the secondary excitation codebook. It is known that the quality of coded speech, especially during voiced regions, is strongly influenced by the continuity of its harmonic structure [2]. For steady voiced regions this continuity is primarily maintained by the LTP. However, the CELP algorithm fails to adequately model transitional regions of speech, such as voiced onsets. This is due to the fact that the LTP memory cannot build up fast enough to track these changes. Consequently, it is left to the secondary excitation codebook to compensate for the LTP's loss of performance. In the case of CELP, the secondary excitation codebook is solely populated with gaussian random vectors which are inadequate for modelling these regions. This drawback is dealt with by PRELP by using a pulse like excitation signal to model the remaining long term correlations. In PRELP the excitation vectors are formed by placing a unit amplitude pulse at the start of the excitation vector $\boldsymbol{x}$, and then every P samples. P is varied from D_{min} (smallest possible pitch) to L (subframe size) to get all primary vectors. Whilst D_{min} is usually related to the minimum pitch, it can be varied to enhance fidelity. In the case of this wideband coder D_{min} is set to 19, typically half the smallest pitch. For each P, the primary candidate excitation vector is derived as follows:

$$x_j(n) = \begin{cases} 1 & n = iP < L, \quad i = 0,1,2\ldots \\ 0 & otherwise \end{cases} \tag{1}$$

For each primary vector $\boldsymbol{x}_j$, P-1 further vectors $\boldsymbol{x}_{j+k}$, $k=1,2,\ldots,P-1$, are derived by shifting the pulses as,

$$x_{j+k}(n) = \begin{cases} 0 & n = 0,1,2,\ldots,k-1 \\ x_j(k-n) & n = k,k+1,\ldots,L-1 \end{cases} \tag{2}$$

in order to form all possible phase positions. Thus the number of candidate excitation vectors C depends on L and D_{min} such that

$$C = L + \sum_{I=D_{\min}}^{L/2} I + \sum_{I=1}^{L/2-1} I \tag{3}$$

The number of bits required to transmit each excitation index is $B=log_2C$. If C does not correspond to an integer power of 2 a further 2^B-C vectors are then searched for in

the fixed codebook. By forcing the secondary excitation to have pitch structure, it is possible to match voiced onsets and other transitional regions of speech more accurately. This is because, firstly, the LTP memory builds up faster to track the incoming periodicity and secondly the secondary excitation provides the required periodicity where the LTP fails. For this application 256 PRELP excitation vectors are searched in total.

In addition to PRELP vectors the secondary excitation also consists of a codebook of 256 overlapping sparsely populated gaussian sequences. The purpose of these sequences is to model those regions of speech whose residual signal contains little or no periodicity. For example, steady state voiced regions where the LTP has a high matching performance, or unvoiced regions where the original signal contains very little periodicity. Each gaussian excitation vector consists of 5 random numbers with an equal spacing of 8 samples between each adjacent pulse. The next excitation vector is formed by shifting each random pulse by one sample. This is repeated 7 times, each shift being treated as a new excitation vector. Once all pulse positions have been exhausted a new primary excitation vector is formed by the overlapping process, and the procedure repeated for the next 7 shifts. By using this method the decoder need only store 36 random numbers to form the codebook, hence keeping the memory requirements to a minimum. Each secondary excitation codebook search requires all 512 entries are searched, and a single vector is chosen based on the weighted Minimum Mean Square (MSE) criterion.

Investigations show that gaussian excitation vectors are selected in preference to PRELP vectors by the ratio of 2:1. This indicates that in two thirds of all cases, the remaining signal which is to be modelled by the secondary codebook is random in nature. In the other third of cases long term correlations are still present. To determine if a particular speech region is modelled by either gaussain or PRELP vectors the following normalised STP and LTP matching criteria are used:

$$M_{stp} = \frac{\sum_{n=0}^{L-1} s_m(n) s_1(n)}{\sqrt{\sum_{n=0}^{L-1} s_m^2(n) \sum_{n=0}^{L-1} s_1^2(n)}} \quad (4)$$

$$M_{ltp} = \frac{\sum_{n=0}^{L-1} s_{ltp}(n) s_2(n)}{\sqrt{\sum_{n=0}^{L-1} s_{ltp}^2(n) \sum_{n=0}^{L-1} s_2^2(n)}} \quad (5)$$

where $s_1(n)$ and $s_2(n)$ are the first (original) and second (STP memory-subtracted) reference speech signals, and $s_m(n)$ and $s_{ltp}(n)$ are the STP and LTP filter memory responses respectively. M_{stp} and M_{ltp} can vary between -1.0 and +1.0. When M_{stp} is negative it is assumed that there is no match, this is due to the fact that the STP filter has no scaling factor. However, the sign for M_{ltp} is ignored as the LTP response consists of a signed gain factor. The regions of speech are characterised according to the following bounds:

(1) $M_{stp} \geq 0.8$, voiced offset where the STP memory makes up most of the output speech.
(2) $M_{ltp} \leq 0.4$, unvoiced speech.
(3) $M_{ltp} \geq 0.9$, steady state voiced speech, high LTP match.
(4) $M_{stp} < 0.8$ and $0.4 < M_{ltp} < 0.9$, undetermined, transitional regions including voiced onsets.

By using the above stated measures we were able to show that if the speech was determined to be in the first three categories then a gaussian random vector was most likely to be chosen. However, if the region of speech falls within the bounds of the fourth category the probability of choosing a PRELP vector was considerably greater than obtained for the previous three. This indicated that PRELP vectors were primarily being selected to model regions of speech where the LTP performance was not as high. For example, transitional regions of speech such as voiced onsets where the residual signal still contained long term correlations.

The wideband coder uses adaptive spectral shaping of the secondary excitation vector. The AbS modelled spectrum of STP and LTP inverse filtered speech is assumed to be flat, which in actual speech is not the case. This is especially applicable to PRELP where a single pulse in the time domain corresponds to a flat frequency response. Adaptive spectral shaping is applied to the secondary excitation to compensate for these model inaccuracies. In wideband speech it is important to adequately represent the higher frequencies as they contribute to the overall perceived quality and brightness of the coded speech. To achieve this desired effect a spectral tilt towards the higher frequencies is added to the secondary excitation vector. Thus, by emphasising these lower energy components we ensure that sufficient coding effort is applied to them. The adaptive spectral shaping should be included in the AbS loop to improve its overall effectiveness. The transfer function of the adaptive shaping filter is therefore given by:

$$H_s(z) = \frac{\left(1 - \sum_{i=1}^{p} a_i \beta^i z^{-i}\right)}{\left(1 - \sum_{i=1}^{p} a_i \alpha^i z^{-i}\right)} \cdot \left(1 - \mu z^{-1}\right) \quad (6)$$

Typical values for α, β and μ are found to be 0.7, 0.9 and 0.3 respectively. A block diagram of the PRELP wideband encoder with enhanced spectral adaptive

shaping is shown in Figure 1. The parameter bit allocation is shown in Table 1.

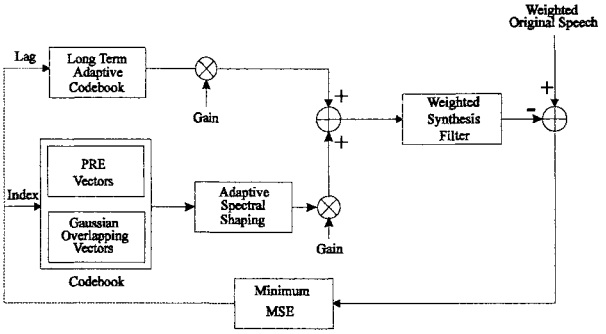

Figure 1 Wideband PRELP Coder

3. PARAMETER QUANTISATION

The LSFs are quantised using 16 independent non uniform scalar quantisers, a total of 60 bits is allocated for this purpose. The pitch gain is quantised using a standard 5-bit non uniform scalar quantiser. The secondary excitation codebook gain is quantised using a novel adaptive technique which is described in the following section.

The number of bits required to quantise the secondary codebook gain is primarily determined by its dynamic range or variance. A further constraint is that the distortion introduced by the process should be imperceptible. Various schemes have been proposed to reduce this gain variance and thus the required bit rate such as the prediction method used in LD-CELP[3]. In this section, an alternative method is outlined where the quantiser is adaptive with respect to the LPC synthesis filter. Initially the secondary codebook gain for each subframe is normalised to the overall RMS energy of the current frame of original speech. The reason for this is that typically a high value of secondary codebook gain is associated with high energy speech and vice versa. Thus, exploiting the relationship between these two parameters results in the desired effect of reducing the dynamic range of the secondary codebook gain. This is evident by comparing the pdfs shown in Figures 2(a) and 2(b).

The resultant gain is then further normalised to the RMS energy response of the MSE selected excitation vector at the output of the LPC synthesis filter. Let $\boldsymbol{x}_{sel}$ be the MSE selected excitation vector from the codebook. Then the synthetic output due to $\boldsymbol{x}_{sel}$ can be expressed as:

$$y_{sel} = \sum_{i=0}^{n} h_w(i) x_{sel}(n-i) \quad 0 \le n \le L-1 \qquad (7)$$

where $h_w(n)$ is the weighted STP impulse response. The RMS energy of the STP synthesis filter response to the MSE selected excitation vector is:

$$RMS_{exsel} = \sqrt{\frac{1}{L}\sum_{n=0}^{L-1} y_{sel}^2(n)} \qquad (8)$$

The residual secondary codebook gain (g_{res}) can be expressed as:

$$g_{res} = \frac{g_c \cdot RMS_{exsel}}{Q\left[RMS_{frame}\right]} \qquad RMS_{exsel} \le 1 \qquad (9)$$

Note that $Q[.]$ denotes the quantisation process. The frame RMS energy should be quantised before the residual secondary codebook gain is calculated.

This results in the normalisation process being adaptive to the gain of the LPC filter and also takes into account the energy variation, at the output of the synthesis filter, which is due to the different positions and number of pulses present in the selected excitation vector from one subframe to the next. For example, excitation vectors whose pulses are concentrated at the beginning of the subframe induce a higher energy at the output of the LPC synthesis filter than those whose pulses are concentrated towards the end. By comparing the pdfs shown in Figures 2(b) and 2(c) it can be seen that this process has the effect of vastly reducing the variance of the codebook gain. This is especially applicable to coders such as PRELP and ACELP[4] whose excitation vectors vary in both the position and number of pulses from one subframe to the next.

Both the RMS energy of the current frame of original speech (RMS_{frame}) and residual codebook gain are quantised using non uniform scalar quantisers. RMS_{frame} is calculated and transmitted on a frame basis, whereas the residual gain operation is performed on a subframe basis. Clearly, RMS_{exsel} is found at the decoder by exciting the LPC filter with the excitation vector which corresponds to the current subframe's received codebook index.

Informal listening tests were used to find the optimal number of bits required to represent the parameters. It was found that there was no perceived distortion due to the quantisation process when 6 bits were used for the overall frame RMS energy and 5 bits for each of the subframe's residual gains. The parameter bit allocation for the coder is shown in Table 1.

Parameters	Bits per Frame	Bit Rate
16 LPC	60	3000
Frame RMS	6	300
8 Pitch Gain	40	2000
8 Pitch Index	64	3200
8 CB Res. Gain	40	2000
8 CB Index	72	3600
Total	282	14100

Table 1: Wideband PRELP coder bit allocation

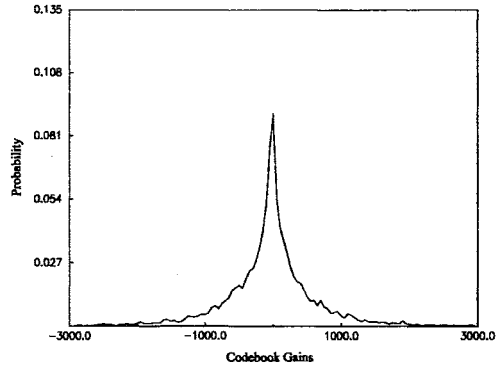

Figure 2(a) pdf of Secondary Codebook Gains

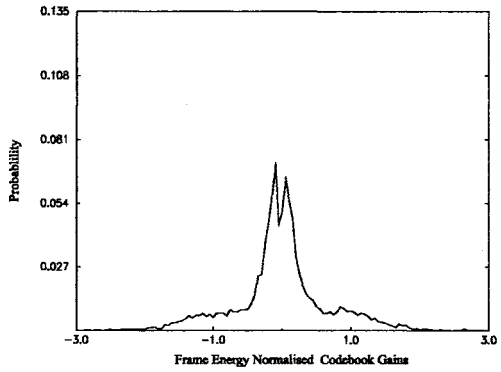

Figure 2(b)pdf of Frame Energy Normalised Codebook Gains

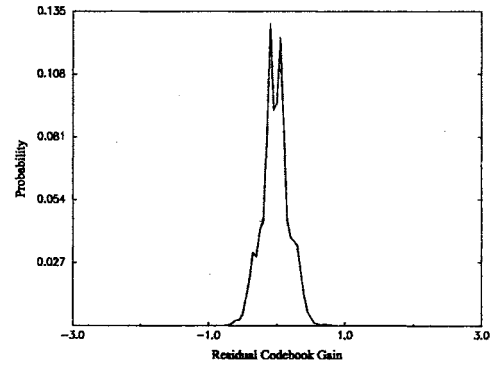

Figure 2(c) pdf of Frame and LPC filter Energy Normalised Codebook Gains

4. CHANNEL ERROR PROTECTION

The channel error performance depends mainly on the way the parameters are quantised and error protected. Due to restriction of available bandwidth and decoder complexity, only the two most error sensitive PRELP parameters were protected, i.e. the LPC coefficients and excitation vector gain.

We protected the LPC coefficients using our previously developed and very successful built in error control scheme. By using the Line Spectrum transformation, the monotonicity and frequency spacing of the LSFs can be used for error detection and correction [5][6]. With this scheme 0.1% errors on the LPC coefficients were not noticeable. For the excitation vector gains it was decided to use Forward Error Correction (FEC). It is well known that FEC decoders can be computationally complex, thus for this application a (7,4) Hamming code is used. It is stated in section 3 that to fully decode the excitation gain at the receiver it is necessary to have three separate parameters: frame RMS energy, residual gain and codebook index. It is noted that the excitation gain is mainly sensitive to channel errors in the frame RMS energy. Thus only the four most significant bits of the frame RMS energy are protected using the three parity bits , the least two significant bits are left unprotected. Therefore for a total FEC redundancy of 150 bits/sec the resultant coder is found to be transparent to random errors up to 0.1%.

5. CONCLUSIONS

In this paper we have presented a wideband speech coder based on the PRELP algorithm operating at a source rate of 14.1kb/s. The coder has been designed to fit three main criteria; low computational complexity, high speech quality and good channel error performance. The decoder is of sufficient low complexity to be implemented on a single fixed point DSP device. This has been carried out using a Motorola 56000. An innovative new method for quantising the secondary codebook gain has resulted in a low FEC redundancy for the protection of this parameter. In recent in-house tests, it was found that the speech quality of our coder was comparable to that produced by the wideband audio standard of G722 operating at 48kb/s.

REFERENCES

[1] M.R.Schroeder, B.S.Atal "Code Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", Proc. of ICASSP-85, pp 937-940.

[2] A.M.Kondoz, J.Horos, B.G.Evans, M.R.Suddle "Pulsed Residual Excited Linear Prediction" IEE Proc.-Vis Image Signal Process., Vol. 142, No. 2, April 1995.

[3] Juin Hwey Chen "High-Quality 16kb/s Speech Coding With a One-Way Delay Less Than 2ms". Int. Proc. of ICASSP-90, pp 453-456.

[4] C. Laflamme et al, "On redcuing computational complexity of codebook search in CELP coder through the use of algebraic codes" Int. Proc. of ICASSP-90, pp177-180.

[5] M.R. Suddle, A.M. Kondoz, B.G. Evans "DSP implementation of Low Bit Rate CELP Based Speech Coders" Proc. Int. Conf. on Digital Processing of Signals in Communications, Loughborough, U.K. Sep 1991, pp 309-314.

[6] S.A. Atungsiri "Joint Source & Channel Coding for Low Bit Rate Commincation Systems, PhD Thesis University of Surrey, Guildford U.K. 1991.