WIDEBAND SPEECH CODER AT 13 Kbit/s

M. Ould-cheikh

ABSTRACT

A 13 kb/s wideband CELP speech encoder was developped. This is an area of increasing growth and interest due to some emerging applications like :multimedia devices, videoconferencing, ISDN applications, etc; these scenarios require high-quality speech without the constraint of the limited telephonic bandwidth. Thus, the bandwidth considered in those applications goes from very low frequencies (around 50 Hz) up to 7000 Hz. The sampling frequency typically used is 16 kHz, although higher sampling frequencies are under consideration for some applications. The research goal consists of reducing the bit rate while maintaining the subjective quality. One way to approach the problem is to extend the telephonic bandwidth schemes to this scenario, tuning them to handle chiefly speech, but also music. The CELP algorithm is used for achieving a toll quality of speech at 13 kb/s. We have introduced a pitch predictor to restore the periodicity of speech signal. In order to reduce the computational complexity, we used an algebraic codebook and the Backward Filtering technique.

KEYWORDS

Speech coding, CELP algorithm, Pitch predictor, Backward filtering.

Teacher-researcher, Dpt. Of Electronics, communications system Laboratory, Polytechnical military School, Algiers, Algeria.

1. INTRODUCTION

In the recent years, there has been a great advance in the development of speech coding algorithms at very low bit rates [1]. High-quality speech coders are now available at bit rates below 8 kb/s. Researchers efforts, however, have focused on narrow-band speech signals where the transmission bandwidth is limited to 300-3400 Hz, as in analog telephone systems. This bandwidth limitation degrades the speech quality, specially when the speech is to be heard through loudspeakers. For many future applications, a wider bandwidth is needed in order to achieve face-to-face communication quality. A bandwidth of 50-7000 Hz was found appropriate, resulting in significantly improved quality as compared to narrow-band speech. The quality improvements are in terms of increased intelligibility, naturalness and speaker recognition. High frequency enhancement (3400-7000 Hz) provides greater intelligibility and fricative differentiation, and low frequency enhancement (50-200 Hz) contributes to increased naturalness [2]. Transmission of low frequencies is made possible by the end-to-end digital connectivity of future communication systems. Several future applications are foreseen for wideband (50-7000 Hz) speech coders, such as teleconferencing, commentary channels, and high-quality wideband telephony.

Code-excited linear predictive (CELP) coding is one of the most promising algorithms to give high speech quality at such low bit rates. The main drawback of the CELP is its excessive computational complexity.

This paper involves the investigation of algebraic CELP (ACELP) coding of wideband speech. The aim is to develop ACELP coding techniques able to produce high quality wideband speech at bit rates below 16 kbit/s..

2. RESEARCH ON WIDEBAND ACELP

The work involves the tuning of of the coder parameters in order to achieve high quality wideband speech. This includes the parameters of the linear prediction (LP) analysis, perceptual weighting, and codebook shaping filters. It also includes the selection of the proper sizes of LP update frames and excitation frames, and the size of the excitation codebook.

As we move from narrowband to wideband, the sampling frequency is doubled, and larger frames sizes are needed to maintain low encoding bit rates. Subsequently, very large excitation codebooks are needed in order to maintain high speech quality. This gives rise to a complexity which is difficult to be handled by the existing CELP algorithms. The approach which is traditionally followed is to split the bandwidth into lower frequency and upper frequency parts and independently encode each band. This, however, requires a fixed allocation of the bit rate between the two bands, which is not the most efficient approach as the amount of energy in the bands differs widely from one frame to another. Due the efficiency to the ACELP algorithm, as we will see later, we were able to follow a full band approach in encoding the wideband speech. There is no fixed division of the bit rate between the different bands, however, the ACELP favors the spectral portions with higher energies, thus the bits are allocated to the perceptually important regions. Preemphasizing the input speech

reduces the spectral dynamic range and ensures that the high frequencies are efficiently encoded [3],[4].

3. LINEAR PREDICTION ANALYSIS AND PITCH ANALYSIS

Linear prediction analysis is performed to obtain the parameters of the synthesis filter, or the short-term predictor. This filter describes the short-time spectral envelope of the speech signals and it is updated every 15-30 ms. The synthesis filter is given by

$$H(z) = \frac{1}{1 - \sum_{k=1}^{m} a_k z^{-k}} \qquad (1)$$

where a_k, k= 1,...,m, are the predictor coefficients, and m is the predictor order. The LP analysis is performed using the autocorrelation method. In this method, the first m+1 autocorrelations of the Hamming-windowed speech are computed, and the LP parameters are determined by solving the Toeplitz system of m equation using the Levinson-Durbin algorithm [5],[6].

In general, the autocorrelation method ensures the stability of the synthesis filter(i.e. the poles of H(z) are inside the unit circle). However, as the sampling is increased to 16 ksample/s in the wideband case, some stability problems are encountered. This higher sampling frequency and the higher filter order needed for wideband speech result in filters with very high prediction gains (or very small prediction errors). Using single precision arithmetic occasionally gives rise to singular matrices in the Levinson-Durbin algorithm as the predictor order is increased. The first is to preemphase the input speech signal. This accomplished by filtering the input speech by the single zero filter $1-\mu^{-1}$. The preemphase has two advantages. It reduces the dynamic range of the input signal resulting in lower required precision, and it emphasizes the higher frequencies in the speech signal, so that higher frequencies can be accounted for by the fixed order model of the ACELP algorithm. The second procedure used to improve the LP analysis is to perform lag windowing on the autocorrelations of speech prior to solving the Toeplitz system of equations [4]. Lag windowing has the effect of widening the bandwidths of the speech formants, thus avoiding bandwidth underestimation which is manifested by extremely sharp peaks in the spectral envelope. A binomial window is used where the autocorrelations are modified by

$$r'(i) = r(i)\exp\left[-\frac{1}{2}\left(\frac{2\ f_0 i}{f_s}\right)^2\right], \qquad i = 1,\ldots,m, \qquad (2)$$

where f_0 is the bandwidth expansion and f_s is the sampling frequency. The first autocorrelation r(0) (the main diagonal in the Toeplitz matrix) is increased by 0.0003%, which is equivalent to adding a noise floor that is 45 dB below the speech power . This alleviates the occasional ill-conditioning of the matrix of autocorrelations. Concerning the filter order, we found that an order of 16 was sufficient to model the

envelope of the short-time speech spectrum. Increasing the order did not result in any significant improvement.

Pitch analysis is performed every 4-7.5 ms, and consists of determining the pitch delay and gain. The pitch synthesis filter models the fine structure of the speech spectrum, and is given by

$$H(z) = \frac{1}{1-\beta z^{-M}} \tag{3}$$

where B is the pitch gain and M is the pitch delay. For human speech, the delay could extend from 2.5 to 20 ms. Female speakers have lower delay periods (higher pitch frequencies). In our case, the delay is in the range 40-295 samples, which is encoded with 8 bits. The pitch parameters are encoded inside the analysis/synthesis loop using an efficient procedure.

A simple but effective pitch predictor is the so-called one tap predictor

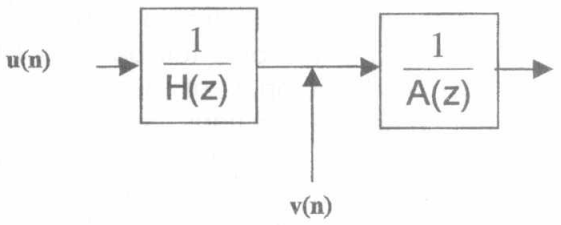

Fig.1.

$$v(n) = u(n) - \beta v(n - M) \tag{4}$$

The pitch parameters β and M are determinated in a closed loop configuration using the analysis by synthesis diagram of figure 2

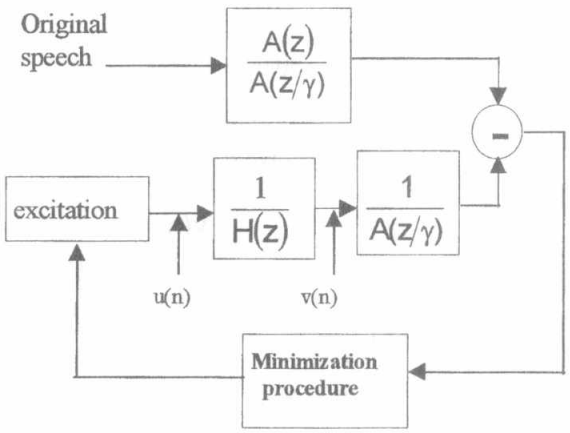

Fig.2.

The advantage of determining the pitch parameters within the analysis loop is that the pitch filter is then optimally contributing to the minimization of the weighted error.

Again we wish to minimize the mean squared error E between the original and synthetic speech signals. One way of reducing the complexity is by obtaining the pitch predictor and the pulse excitation sequentially in two steps.

First we assume that the pulse excitation is zero and search for delay value M and the predictor gain β such that E is minimized. Next, the pitch predictor is held constant and pulse excitation is computed.

4. EXCITATION CODEBOOK

The excitation signal needed to drive the pitch synthesis and LP synthesis filters is selected from a large codebook of innovation sequences by minimizing the perceptually weighted error between the original and synthesized speech. The excitation frame range from 3.75 to 5 ms (60-80 samples). Larger frames would require huge codebook sizes. The ACELP uses an efficient algebraic codebook which could be as large as 2^{20} entries. The codebook need not be stored, and it is very efficiently searched using a focussed search strategy. This ability to incorporate such huge codebook sizes permitted the use of a full band ACELP approach. An excitation codeword contains a small number of nonzero pulses, with fixed amplitudes (-1 or +1) and predefined sets of positions. Searching the codebook is in effect finding the positions of each nonzero pulse which minimize the error criterion. Thus the selected excitation codeword will have, to some extent, optimum pulse positions, within the constraint of fixed amplitudes and limited set of positions for each pulse.

The codebook excitation is computed by minimizing the weighted mean squared error between the speech signal s(n) and the synthetic speech signal $\hat{s}(n)$. Since the weighted filter is linear, the minimization is equivalent to minimizing the unweighted speech signal y(n) and the corresponding weighted synthetic signal $\hat{y}(n)$ as schown in figure 2

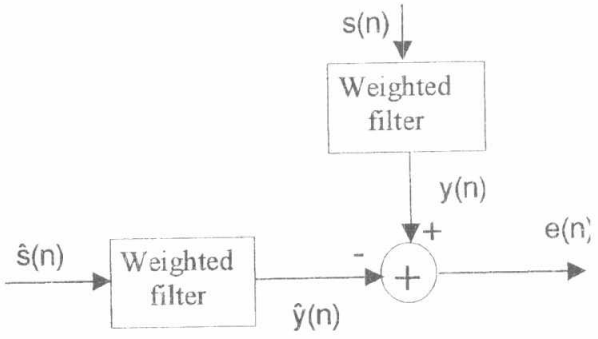

Fig.3.

The speech signal is filtered through the weighted filter $W(z)=\frac{A(z)}{A(z/\gamma)}$ to obtain y(n)

$$y(n)=s(n)+\sum_{k=1}^{p}a_k s(n-k)-\sum_{k=1}^{p}a_k \gamma^k y(n-k) \tag{5}$$

The weighted synthetic speech signal $\hat{y}(n)$ is generated by convolving the excitation signal u(n) with the truncated impulse response h(n) of the cascaded long-term, short-term and weighted filters. Then

$$\hat{y}(n)=\sum_{i=0}^{p}h(i)u(n-k)+\hat{y}_0(n) \qquad 0<n<N \tag{6}$$

where $\hat{y}_0(n)$ is the contribution to $\hat{y}(n)$ from the filter memory.

The weighted error signal energy is expressed as:

$$E=\|x-gHc_k\|^2 \tag{7}$$

where x is the target vector given by the weighted input speech after substracting the zero-input response of the weighted synthesis filter $1/A(z/\gamma)$, g is a scaling gain factor, and H is a lower triangular matrix constructed from the impulse response of the weighted synthesis filter.

The optimal innovation sequence is the codeword which minimizes the distortion measure E_k.
Setting $\partial E/\partial g = 0$ in Equation (7) yields

$$g=\frac{\left(x^T Hc_k\right)}{\|Hc_k\|^2} \tag{8}$$

and substituting Equation (8) in (7) gives

$$E_k=\|x\|^2-\frac{\left(x^T Hc_k\right)^2}{\|Hc_k\|^2} \tag{9}$$

The optimum codeword c_k is selected by maximizing the following quantity as

$$\max_k \left|\frac{\left(x^T Hc_k\right)}{\|Hc_k\|}\right| \tag{10}$$

This codeword search procedure requires a lot of calculations, since each codeword must be filtered at least once per LPC analysis frame. Several alternate approaches have been proposed for reducing the computational complexity. We will

use the backward filtering technique [7], which consists of rewriting expression (10) as

$$\max_k \left| \frac{(x^t H) c_k}{\alpha_k} \right| \qquad (11)$$

where $\alpha_k = \sqrt{\|Hc_k\|^2}$ (the term "backward filtering" comes from the interpretation of (**XH**) as the filtering of time-reversed **X**). The term α^2, represents the energy of the filtered codeword. Because of this energy term, the backward filtering technique does not, of itself, reduce computational complexity when using a stochastic codebook.

The codeword c_k is generated and then filtered in the decoder using the cascaded long-term ant short-term LPC synthesis filters to produce the desired synthetic speech signal.

5 CODER/DECODER SYNOPTIC

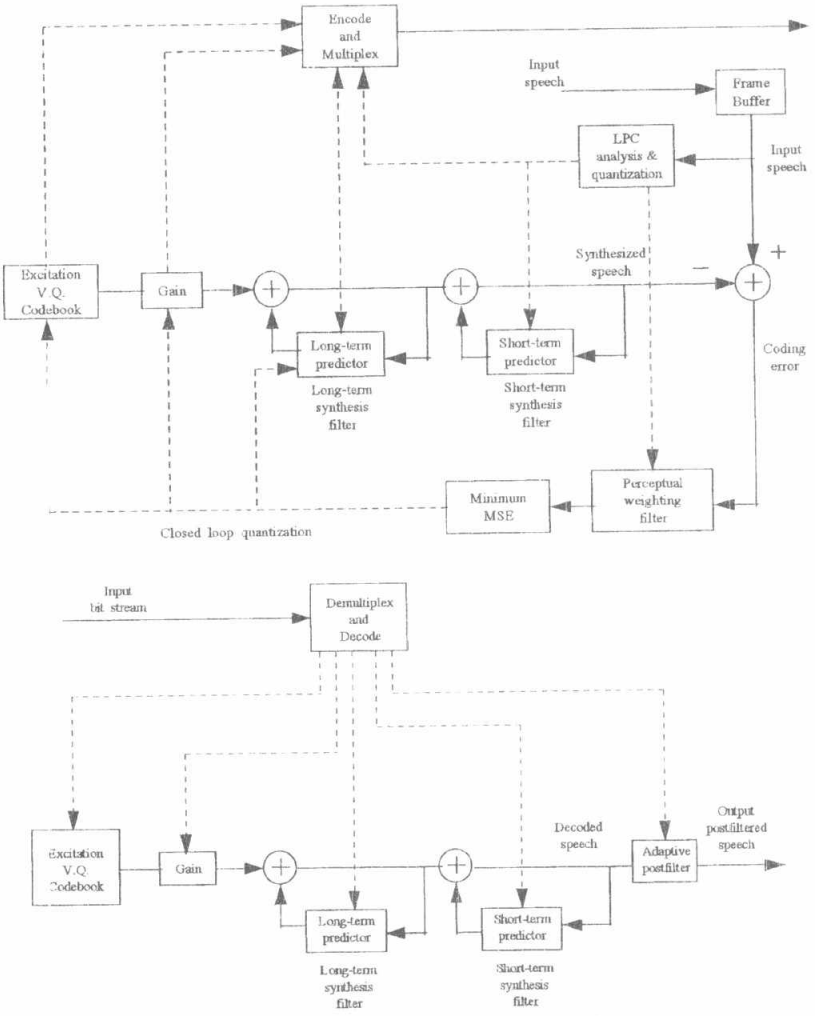

Fig.4.Coder/decoder synoptic

7 12.6 Kb/s HIGH QUALITY WIDEBAND ACELP

In this section we describe the bit allocation of a wideband ACELP coder at 12.6 kb/s which results in high quality encoded speech. The bit allocation is given in Table 1.

Parameter	Update interval (ms)	Bit no
LP filter	15	53
Pitch delay	3.75	8
Pitch gain	3.75	4
Codebook index	3.75	16
Codebook gain	3.75	6

Table 1: Bit allocation for 12.6 kb/s ACELP coding

The speech frame is of 15 ms size (240 samples). The LP parameters are computed using a 30 ms Hamming window centered at the end of the frame. A 16 filter order is used, and the coefficients are quantized using the line-spectral frequency (LSF) representation. The LSFs are quantized with 53 bits ($f_2,...,f_6$ each with 4 bits and the rest each with 3bits). The quantization is performed using a backward adaptive approach [8] where the ordering property of the LSFs is efficiently used to reduce the range of the quantization interval. The information used in quantization involves the limits of each LSF and the limits of the differences between adjacent LSFs. In designing the quantizer, statistics where carried using a speech database of 5 minutes of sentences uttered by the male and female speakers in both French and English languages.

The 15 ms speech frame is divided into 4 subframes of 60 samples (3.75 ms). The pitch and excitation parameters are updated every subframe. The pitch gain is restricted to the range 0-1.2 and quantized with 4 bits. The 60 samples excitation vectors contain 4 nonzero pulses where each pulse can take 16 possible positions (4 bits). Thus, 16 bits are needed to encode the address of the excitation vector (2^{16} codebook). The positions of the pulses are given in table 2. The magnitude of the excitation gain is logarithmically quantized with 5 bits and 1 bit is used for the sign. In a 15 ms speech frame, 189 bits are transmitted, which results in 12.6 kb/s coding.

Pulse no	Amplitude	Set of positions
0	+1	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56
1	-1	1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, 45, 49, 53, 57
2	+1	2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58
3	-1	3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, 47, 51, 55, 59

Table 2: Amplitudes and positions of the 4 pulses in the 2^{16} sized codebook

7.1. Phrases test

Female speakers

P1 : I must have reread that article three times before I realized what was bothering me.

P2 : Quand il s'est réveillé, il était trop tard. Huit satellites ont été mobilisés.

P3 : Là bas il y a de mauvaises vagues très hautes. C'est la question que tout le monde se pose.

Male speakers

P4 : La voiture s'est arrêtée au feu rouge. La vaisselle propre est mise sur l'évier.

P5 : The other memorable event in that conference was the worst presentation I have ever heard.

P6 : Je ne peux atteindre les bocaux de confiture. Dans cette crèmerie on vend du fromage fort.

In order to evaluate the performance of the coder, we have computing the segmental SNR for the sequences of speech. The results are listed in Table 3

Phrase	SNR(dB)	SNRSEG (dB)
1	12.64	8.93
2	11.97	10.75
3	12.36	9.05
4	10.34	7.72
5	9.65	9.53
6	10.19	8.23

Table.3. SNR &SNRSEG for different phrases

8 CONCLUSION

To conclude, the ACELP coding techniques have been successfully adapted to encode wideband speech signals at bit rates below 13 kb/s, where high quality speech was obtained. The preliminary listening tests suggested a high quality encoded speech at 12.6 kb/s. The algorithm was also successful in encoding specific music signals (speech like signals), although it was not as promising with some other signals

REFERENCES

[1] B.S.Atal et al. *Advances in Speech Coding*, Kluwers Academic Pub., 1991

[2] N.S.Jayant, " High-qulity coding of telephone speech and wideband audio," IEEE Communications Magazine, pp. 10-20, Jan. 1990

[3] A. Fuldseth, E. Harbourg, F.T. Johansen. J.E. Knudsen, "Wideband speech coding at 16 kbit/s for videophone application", Speech communication, vol. 11, pp 139-148, 1992.

[4] C. McElroy, B.P. Murray and A.D.Fagan," On improving wideband CELP speech coders", Proc, EUSIPCO-94, Signal Processing VII: Theories and applications, Elsevier Science Publishers, 1994,pp. 912-915

[5] J.D.Markel and A.H.Gray, Jr., *Linear Prediction of Speech,*Springler-Verlag, 1976.

[6] Y.Tohkura and F.Itakura, " Spectral smoothing technique in PARCOR speech analysis-synthesis," IEEE Trans. On ASSP, vol. 26 no. 6, pp 587-596, Dec. 1978.

[7] C. Laflamme, J.P. Adoul, H. Y. Su and S. Morissette, "On reducing computational complexity of codebook search in CELP coder through the use of Algebraic codes" ICASSP june 1991.

[8] N.Sugamura and N.Farvardin, "Quantizer design in LSP analysis-synthesis," IEEE j. on Selec. Areas in Commun., vol. 6, no. 2, pp. 432-440, Feb. 1988.