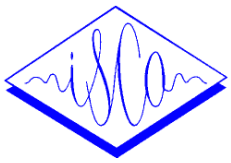

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

IMPROVING THE SPEECH QUALITY OF CELP-CODERS BY OPTMIZING THE LONG-TERM DELAY DETERMINATION

Balss, U., Kipper, U., Reininger, H., and Wolf, D.

Institut für Angewandte Physik der Universität Frankfurt a. M.
Robert-Mayer-Str. 2-4, 6000 Frankfurt am Main, FRG

ABSTRACT

In this contribution we report on a method to control the delay determination in CELP coders, which improves the subjective speech quality without increasing the bit rate. Possible values of the delay of the long-term predictor were scored with respect to previous delay values. This scoring is done by weighting the closed-loop performance of the long-term predictor by means of a weighting factor, which depends on the difference between the previous delay value and the actual delay value. The final decision on the delay value is made on the basis of the weighted performances. Optimization and evaluation of this method were studied in simulation experiments using conventional CELP with stochastic excitation as well as ACELP with adaptive excitation at low data rates [2,3]. CELP coders using the new delay determination achieve a significant improvement of subjective speech quality. With an ACELP coder a good speech quality can be obtained even at very low data rates.

Keywords: ACELP, CELP, Long-Term Prediction.

1. INTRODUCTION

Code-Excited-Linear-Prediction (CELP) is a very efficient method for encoding telephone bandwidth speech at low data rates [10]. Natural sounding speech with acceptable quality can be obtained at bit rates around 4.8 kb/s. A certain kind of roughness is still noticable, especially in voiced parts of the speech signal. This is mainly due to restrictions of the long-term predictor parameters [9]. For realizing data rates below 6 kb/s the adaptation length of the long-term predictor has to include 40 samples or more. The range of possible delay values is limited, mostly in the range of 32-160. The long-term predictor is only of first order and the predictor coefficient is coarsly quantized.

To investigate the problems arising from these restrictions a detailed analysis of the long-term predictor was made with recordings of isolated vowels spoken by a professional speaker with the task not to change his pitch period during the recording. Fig. 1 shows the delay k_p resulting from the closed-loop optimization technique in a 4.8 kb/s CELP coder with the vowel /u/ as input speech signal. In this speech signal the pitch period corresponds to k_p=60. However, the delay is alternating between k_p=60 and k_p=120. Due to this delay fluctuations the reconstructed speech signal sounds very rough. Restricting the delay k_p to a range of $55<k_p<65$ as shown in Fig. 2, the reconstructed speech sounds clear without any roughness.

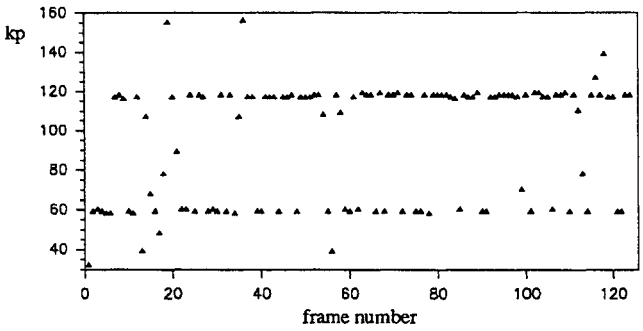

Fig. 1: Delay k_p for the vowel /u/ (128 frames = 2s)

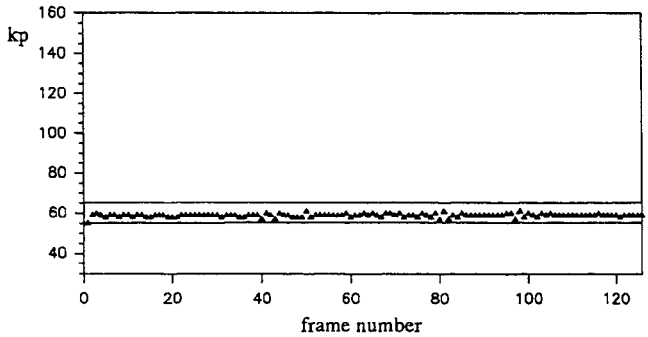

Fig. 2: Delay k_p for the vowel /u/ with limited range of k_p-values (128 frames = 2s)

In order to avoid pitch incompatible delay variations, an algorithms was suggested, which allows to determine non-integer delay values [5]. With this technique effects like those decribed above can be avoided, but additional data rate

for encoding the non-integer delay values is required and the complexity of the scheme is increased drastically. The method for obtaining a smooth delay contour here proposed is based on a modification of the optimization criteria used for the delay determination. Thus, the data rate is not affected and the complexity of the encoding increases only slightly.

2. MODIFIED LONG-TERM DELAY OPTIMIZATION CRITERIA

The general form of a long-term predictor of order P_L is given by

$$A_L(z) = 1 - \sum_{k=-[(P_L-1)/2]}^{(P_L-1)/2} b_k z^{-(k_p+k)} , \qquad (1)$$

with the delay k_p and the filter coefficients $\underline{b}$. In Analysis-by-Synthesis schemes these parameters are determined by minimizing the coding error

$$\varepsilon(\underline{b}, k_p) = \sigma_{s'}^2 - 2\underline{b}^T \underline{r}_{k_p} + \underline{b}^T \underline{\underline{R}}_{\bar{s}\bar{s}} \underline{b} , \qquad (2)$$

with

$$\begin{aligned} \sigma_{s'}^2 &= \sum_{n=1}^{N_S-1} s_W'^2(n) \\ r_{k_p}(j) &= \sum_{n=1}^{N_S-1} s_W'(n)\bar{s}_{k_p}^-(n+j) \\ R_{\bar{s}\bar{s}}(j,k) &= \sum_{n=1}^{N_S-1} \bar{s}_{k_p}^-(n+j)\bar{s}_{k_p}^-(n+k) , \end{aligned} \qquad (3)$$

where $\{s_W'(n)\}$ denotes the input speech signal weighted with $W(z)=A(z)/A(z/\gamma=0.8)$, $\{\bar{s}_{k_p}(n)\}$ the long-term predictor signal, and N_S the adaptation interval length. From

$$\frac{\partial\varepsilon(\underline{b}, k_p)}{\partial b_i^*} = 0 , \; i=-[(P_L-1)/2], \ldots, (P_L-1)/2 , \qquad (4)$$

together with (2), the system of linear equations

$$\underline{\underline{R}}_{\bar{s}\bar{s}} \underline{b} = \underline{r}_{k_p} \qquad (5)$$

results for calculating the optimum filter coefficients $\underline{b}^*$. Inserting the optimum coefficients $\underline{b}^*$ in (1) leads to the coding error

$$\varepsilon(k_p) = \sigma_{s'}^2 - 2\underline{b}^{*T} \underline{r}_{k_p} , \qquad (6)$$

which is a function of k_p only. Finally, the optimum delay k_p^* is given by

$$\begin{aligned} k_p^* &= \arg \min_{k_p} \varepsilon(k_p) \\ &= \arg \max_{k_p} \underline{b}^{*T} \underline{r}_{k_p} \end{aligned} \qquad (7)$$

In order to smooth the delay contour without affecting the dynamics of the k_p-values, necessary for obtaining natural sounding speech, a weighting of the coding error was introduced. Let $k_p(0)$ denote the delay value of an actual frame and $k_p(-t)$, $t>0$, that of the t-th past frame. If the condition

$$D = \sum_{t=1}^{T} \left| k_p(-t) - k_p(-t-1) \right| < \xi \quad , \xi > 0, \qquad (8)$$

is satisfied the coding error of all $k_p(0)$-candidates in the interval $[k_p(-1)-\delta, k_p(-1)+\delta]$ is defined by

$$\varepsilon_w(k_p(0)) = \sigma_{s'}^2 - 2\underline{b}^{*T} \underline{r}_{k_p} \cdot \rho \quad , \rho > 1. \qquad (9)$$

For all other $k_p(0)$-candidates the coding error $\varepsilon(k_p)$ is given by (6). The parameter T determines the number of previous k_p-values which influence the calculation of an actual one. ξ allows to control the smoothness of the k_p-contour. With the parameter δ the range of $k_p(0)$-candidates which are perferred in the delay determination can be adjusted. The factor ρ controls the amount of preference.

3. SIMULATION EXPERIMENTS AND RESULTS

A 4.8 kb/s ACELP coder was used to optimize the parameter configuration of the delay determination procedure [4]. In contrast to a conventional CELP scheme, where the only speech specific information in the excitation arises from the memory of the long-term predictor, in a ACELP scheme the fixed stochastic codebook is replaced by a codebook with speech adaptive content. The adaptive codebook consists of a set of basic excitation vectors, which are adapted to an analysis speech frame by calculating optimum amplitude values. Each basic excitation vector consists of regularly spaced pulses, denoted as pulse grid [6]. The grids are grouped into G different grid classes each comprising all grids with the same number N_p of pulses but different positions of the first pulse. Assuming that every position of an excitation vector with N_S elements occurs as a pulse position in a grid class, the number N_G of pulse grids within a class is defined by N_S and N_p. This allows a very efficient encoding of the pulse positions. Using several pulse grids for encoding consecutive speech frames, excitation sequences with variable pulse rates are realized. The optimum pulse amplitudes with respect to a minimum coding error of all pulse grids are calculated by solving a set of linear equations. The speech signal which has to be encoded is used to define the coefficient matrix of the equation system. The resulting amplitudes are quantized and encoded at a very low bit rate using a gain shape vector quantizer [7].

The parameter configuration of an ACELP scheme for a data rate of about 4.8 kb/s is summarized in Table 1 [8]. An adaptation length of 48 samples was chosen to realize $G=10$

different grid classes with 1 up to 48 pulses per excitation vector. With this large variety of speech adaptable excitations it is possible to achieve a substancially higher speech quality than with stochastic excitations.

Table 1: Bit allocation of a 4.8 kb/s ACELP scheme

Length of frame	192 samples		
Length of subframe	48 samples		
Short-term predictor	LSF, order 10, VQ 24 bit		
Long-term predictor	order 1, coeff. 3 bit, delay 32-160		
Excitation			
Class	Number of pulses N_p / bit	VQ amplitudes	Sum of bits
1	1 / 6 bit	5 bit	11
2	2 / 5 bit	6 bit	11
3	3 / 4 bit	7 bit	11
4	4 / 4 bit	7 bit	11
5	6 / 3 bit	8 bit	11
6	8 / 3 bit	8 bit	11
7	12 / 2 bit	9 bit	11
8	16 / 2 bit	9 bit	11
9	24 / 1 bit	10 bit	11
10	48 / 0 bit	11 bit	11

The delay k_p determined with the unmodified coding error for the speech signal of Fig. 4a with 2s in duration is shown in Figure 4b. It can be seen that the k_p-values of consecutive frames vary strongly, even in voiced speech segments. Efforts to smooth the delay contour by resticting the range of possible k_p-values in (7) depending on the previous optimum k_p-value led to a drastical decrease of the performance of the long-term predictor. Obviously, a good speech quality can only be obtained if a sufficient variability of k_p is provided. Already by switching off the long-term predictor if the prediction gain is less than 0.6 dB a small but noticable improvement in speech quality results.

The parameters of the improved long-term delay determination procedure were optimized in informal listening tests. The best subjective speech qualtity was achieved with ρ=1.8, T=4, ξ=100, and δ=10. With these parameter values the delay contour for the vowel /u/ obtained with the modified optimization criteria is similar to that of Fig. 1b.

The k_p-contour with from the encoding of the speech signal in Fig. 4a with modified error criteria is shown in Fig. 4c. Comparing the k_p-contour with that obtained without weighting (Fig. 4b) one can see the influence of the modified optimization criteria. In particular, in voiced speech segments the delay determination with weighting avoids the strong fluctuation of the k_p-values. The resulting change of the k_p-values in other speech parts had no influence on the performance of the long-term predictor. The difference of consecutive k_p-values determined with weighting is much smaller in voiced parts of the speech signal than in unvoiced

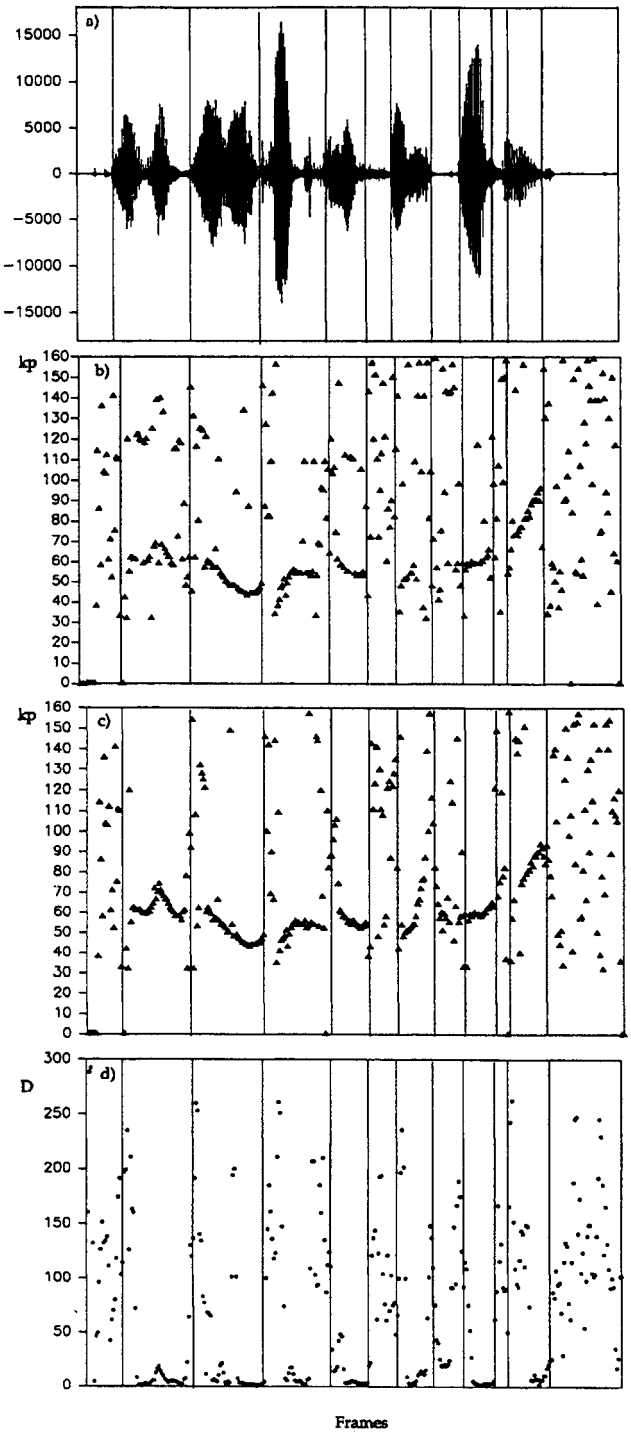

Fig. 4: a) Original speech signal
b) Delay contour
c) Smoothed delay contour
d) Accumulated difference D

parts, as can be seen from Fig. 4d. Therefore, the choice of the parameter ξ in (8) is not very critical. ξ can be chosen in the range of about 40 to 100 without affecting the speech quality.

In the processed speech the roughness due to the k_p-fluctuations is completely eliminated. Further experiments have shown that the speech quality of the 4.8 kb/s ACELP coder with improved long-term delay determination is nearly transparent.

4. CONCLUSION

By modifying the optimization criteria for the determination of the delay value of a long-term predictor in a CELP coder a smooth delay contour in voiced speech segments was obtained. Applying this method for delay determination a significant improvement of subjective speech quality results without increasing the data rate of 4.8 kb/s. The method can improve the performance of a CELP scheme at still lower bit rates.

5. REFERENCES

[1] Gerson, I., Jasiuk, M.: "Vector Sum Excited Linear Prediction (VSELP)", IEEE Workshop on Speech Coding for Telecommunications, 1989, pp. 66-68.

[2] Kipper, U., Reininger, H., Wolf, D.: "Optimization and Efficient Encoding of Excitation Pulses in Low-Bit-Rate MPLPC Schemes", Proc. URSI-ISSSE, 1989, pp. 824-827.

[3] Kipper, U., Reininger, H., Wolf, D.: "Improved CELP Coding Using Adaptive Excitation Codebooks", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1991, pp. 237-240.

[4] Kipper, U., Reininger, H., Wolf, D.: "Improved CELP Using a Fully Adaptive Excitation Codebook", NATO Advanced Study Institute, Bubión 1993, to be published.

[5] Kroon, P., Atal, B.S.: "Pitch Predictors with High Temporal Resolution", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1990, pp. 661-664.

[6] Kroon, P., Deprettere E.F., and Sluyter, R.J.: "Regular-pulse excitation: A novel approach to effective and efficient multipulse coding of speech", IEEE Trans., Acoust., Speech, Signal Processing, vol. ASSP-34, 1986, pp. 1054-1063.

[7] Linde, Y., Buzo, A., and Gray, R.M.: "An Algorithm for Vector Quantizer Design", IEEE Trans. Comm., vol. COM-28, 1980, pp. 84-95.

[8] Paliwal, K.K., Atal, B.S.: "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1991, pp. 661-664.

[9] Ramachandran, R.P., Kabal, P.: "Pitch prediction filters in speech coding", Proc. IEEE Trans. Acoust., Speech, Signal Processing ASSP-37, 1989, pp 467-478.

[10] Schroeder, M.R., Atal, B.S.: "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", 1985, pp. 937-940.