IMPROVEMENTS TO THE ANALYSIS-BY-SYNTHESIS LOOP IN CELP CODECS

J.P. Woodard and L. Hanzo

University of Southampton, UK

Abstract

In this paper extensions to the Analysis-by-Synthesis (AbS) loop used in Code Excited Linear Predictive (CELP) speech codecs are considered. Methods for updating the short-term synthesis filter once the excitation parameters have been determined are examined. We show that significant improvements can be achieved by updating the synthesis filter, similar to those obtained using the well known methods of interpolation and bandwidth expansion. However our proposed method of update avoids the increase in the delay of a codec that is usually associated with interpolation. Furthermore the traditional sequential method of determining the adaptive and fixed codebook parameters is examined and compared to an exhaustive search of both codebooks. Three sub-optimum techniques are proposed for improving the performance of the codebook search while maintaining a reasonable level of complexity. The most complex of these increases the codec complexity by only about 40% but provides 80% of the maximum possible 1.1 dB segmental SNR improvement associated with an exhaustive codebook search.

1 Introduction

In this work we have studied ways of improving the Analysis-by-Synthesis (AbS) structure used in Code Excited Linear Predictive (CELP) [1] speech codecs. The block diagram of the encoder of such a codec is shown in Figure 1. The excitation signal $u(n)$ is given by the sum of a scaled adaptive codebook signal (which adds long-term periodicities during voiced speech) and a scaled signal from a large fixed codebook. This excitation is used to drive a synthesis filter which models the effects of the vocal tract. At the decoder the excitation signal is passed through the synthesis filter to produce the reconstructed speech signal $\hat{s}(n)$. Typically the filter parameters are determined first and then the codebook indices α and k as well as the gains G_1 and G_2 are found. The codebook parameters are chosen to minimise the weighted error between the reconstructed and the original speech signals. In effect each

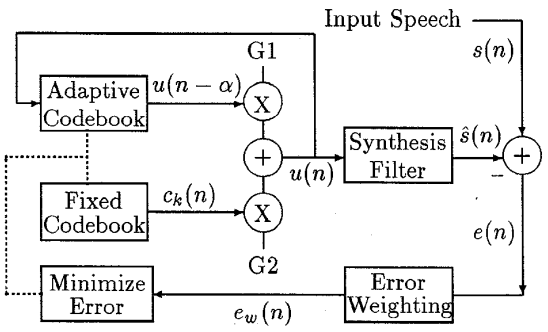

Figure 1: Speech Encoder Schematic

possible codebook entry is passed through the synthesis filter to test which gives an output closest to the input speech in the perceptually weighted sense. This largely closed-loop structure is used in order to produce a reconstructed signal which is as close as possible to the original speech.

There are however two exceptions to a strict closed-loop approach in most CELP codecs. The first is in the determination of the synthesis filter, which is simply assumed to be the inverse of the short-term linear prediction error filter minimising the energy of the prediction residual error. This means that although the excitation signal $u(n)$ is derived taking into account the form of the synthesis filter, no account is taken of the form of the excitation signal when the synthesis filter parameters are determined. This is a deficiency, and means for example that the synthesis filter may attempt to take account of long-term periodicities which would be better left to the adaptive codebook to model.

The second departure from a strict closed-loop approach is in the determination of the codebook parameters. Rather than the adaptive and fixed codebook parameters being determined together to produce an overall minimum in the weighted error signal, the adaptive codebook delay and gain are determined first by assuming that the fixed codebook signal is zero. Then, given the adaptive codebook signal, the fixed codebook parameters are found. This sub-optimum approach is adopted in order to reduce the complexity of CELP codecs to a reasonable level. However it is obvious that it must lead to some degradation in the reconstructed speech.

Radio Receivers and Associated Systems, 26-28 September 1995, Conference Publication No 415 © IEE 1995

In this work we have examined the degradations that result from the two exceptions to the closed-loop approach described above. Furthermore we suggest various algorithms that improve the quality of the reconstructed speech while maintaining a reasonable level of complexity. In our simulations we have used a 4.8 kbits/s Algebraic CELP (ACELP) [2] codec, with the synthesis filter parameters determined every 30ms frame, and the excitation parameters determined every 7.5ms sub-frame.

2 Calculation of the Synthesis Filter Parameters

As described above the synthesis filter is usually simply assumed to be the inverse of the prediction error filter $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} \cdots a_p z^{-p}$ which minimises the energy of the prediction residual for the input speech signal $s(n)$. Here p is the order of the filter, which we took to be equal to ten. It is well known that this is not the ideal way to determine the synthesis filter parameters. Once the excitation signal $u(n)$ has been determined it is possible to re-calculate the synthesis filter coefficients in order to maximise the SNR of the reconstructed speech [3, 4, 5]. In this section we discuss various methods of carrying out this optimization.

We started our investigation of the effects of updating the synthesis filter parameters by finding an upper limit to the improvement possible. The filter coefficients were converted to Line Spectrum Frequencies (LSFs) [6] for quantization, and we used the technique of simulated annealing [7] to find the optimum set of quantized LSFs for each speech frame. Simulated annealing is not a practical method for use in real codecs because of it's complexity. It does however give us an idea of the improvement that can be obtained by updating the synthesis filter parameters, and we found that an improvement of just over 1dB in the segmental SNR of our 4.8 kbits/s codec was possible. With this in mind we attempted to find a method of updating the LSFs that gave a similar improvement without overly increasing the complexity of the codec.

One method of re-optimization which has been tried in conjunction with Multi-Pulse Excited codecs [3, 8] is a Least Squares update. Given an excitation signal $u(n)$ and a set of filter coefficients a_k, $k = 1, 2 \cdots p$, the reconstructed speech signal $\hat{s}(n)$ will be given by

$$\hat{s}(n) = u(n) + \sum_{k=1}^{p} a_k \hat{s}(n-k). \qquad (1)$$

We wish to minimise E, the energy of the error signal $e(n) = s(n) - \hat{s}(n)$ over the frame length L. E is given by

$$\begin{aligned} E &= \sum_{n=0}^{L-1} (s(n) - \hat{s}(n))^2 \\ &= \sum_{n=0}^{L-1} \left(s(n) - u(n) - \sum_{k=1}^{p} a_k \hat{s}(n-k) \right)^2 . \qquad (2) \end{aligned}$$

The problem with Equation 2 is that E is given in terms of not only the filter coefficients but also the reconstructed speech signal $\hat{s}(n)$ which of course also depends on the filter coefficients. Therefore we cannot simply set the partial derivatives $\partial E / \partial a_i$ to zero and obtain a set of p simultaneous linear equations for the optimal set of coefficients.

In the Least Squares approach we make the approximation [8]

$$\hat{s}(n-k) \approx s(n-k) \qquad (3)$$

in Equation 2, which then gives

$$E \approx \sum_{n=0}^{L-1} \left(s(n) - u(n) - \sum_{k=1}^{p} a_k s(n-k) \right)^2 . \qquad (4)$$

We can then set the partial derivatives $\partial E / \partial a_i$ to zero for $i = 1, 2 \cdots p$ to obtain a set of p simultaneous linear equations, which can be solved to give the updated filter coefficients. We invoked this method of update for our 4.8 kbits/s ACELP codec but found that the updated filter coefficients were, in terms of the SNR of the reconstructed speech, usually worse than the original coefficients. This is because of the inaccuracy of the approximation in Equation 3. To obtain any improvement in the segmental SNR of the reconstructed speech it was necessary in each frame to find the output of the synthesis filter using both the original and updated filter coefficients, and transmit the set of coefficients which gave the best SNR for that frame. Using this technique we found that the updated filter coefficients were better than the original coefficients in only about 15% of the frames, and the segmental SNR of the codec was improved by about 0.25dB.

These results were rather disappointing, so we set out to find an improved method of updating the synthesis filter parameters. In recent years relatively new techniques called Total Least Squares [9] and Data Least Squares [10] have been applied to several similar problems, see for instance [11]. We tried these techniques, but found that they were not useful in our situation because a very large number (about 95%) of the sets of filter coefficients they gave resulted in unstable synthesis filters.

In [4] after the initial set of quantized LSFs have been found, a total of 1296 other nearby LSF sets are tried in conjunction with the given excitation. However finding the reconstructed speech $\hat{s}(n)$ and

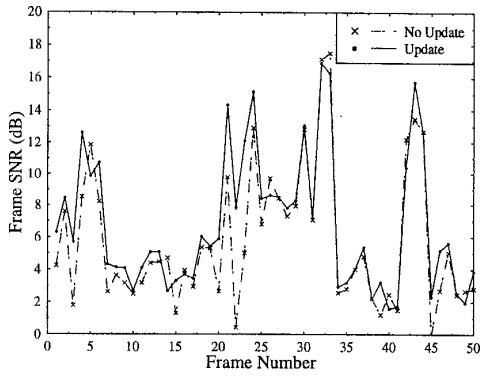

Figure 2: Effect of Update on Variation of SNR

calculating the error energy E such a large number of times gives a very considerable increase in the complexity of the codec. Hence we attempted to employ an alternative technique described below. Starting with the initial set of quantized LSFs we cycle sequentially through all p LSFs in turn, moving them up and then down one quantizer level to see if we can reduce the error energy. Any change which reduces the error energy is accepted. This process can be repeated any number of times, with every testing of all p LSFs counting as one iteration. We found that the gain of this method of updating the quantized synthesis filter parameters saturated after about three iterations, and gave an improvement in the segmental SNR of our codec of just under 1dB. This is almost equal to the improvement produced by simulated annealing of the LSFs, and yet at most only 60 sets of quantized LSFs are tested, while the complexity of the codec is increased by only about 80%.

Not only does updating the synthesis filter help to increase the average segmental SNR of our codec but, as mentioned in [8], it also helps remove the very low minima in SNR that occur for some frames. This effect is shown in Figure 2 which shows the variation of SNR for a sequence of fifty frames for 4.8kbits/s codecs with and without update of the synthesis filter. It is clear that the number of frames with a very low SNR has been reduced by the update. These low minima can be subjectively annoying and so it is beneficial if they can be partially removed.

The results above show that updating the synthesis filter parameters can improve the segmental SNR of our codec by about 1dB, and help remove subjectively annoying low minima in the frame by frame SNR of the codec. Similar improvements can also be achieved using the well known techniques of bandwidth expansion [12] and LSF interpolation [6]. A codec incorporating such techniques is about 10% less complex than our codec using the iterative update procedure described above. However interpolation typically increases the delay of the codec. Our scheme of updating the LSFs provides an alternative which gives similar results and is only slightly more complex, but does not require the delay of the codec to be increased.

3 Calculation of the Excitation Parameters

In order to reduce the complexity of the encoder the error weighting filter in Figure 1 is usually moved so that the input and the reconstructed speech signals $s(n)$ and $\hat{s}(n)$ are separately weighted before their difference is found. For an all-pole synthesis filter of the form $H(z) = 1/A(z)$, where $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} \cdots a_p z^{-p}$, and an error weighting filter $A(z)/A(z/\gamma)$ where γ is a constant set equal to 0.9 in our simulations, the cascade of the synthesis filter and the error weighting filter is equivalent to using a weighted synthesis filter of the form $1/A(z/\gamma)$. The weighted error $e_w(n)$ is then given by

$$\begin{aligned} e_w(n) &= s_w(n) - \hat{s}_w(n) \qquad (5) \\ &= s_w(n) - \hat{s}_o(n) - G_1[u(n-\alpha) * h(n)] \\ &\quad -G_2[c_k(n) * h(n)] \end{aligned}$$

where $s_w(n)$ is the weighted input speech, $\hat{s}_o(n)$ is the zero-input response of the weighted synthesis filter due to it's input in previous sub-frames, and $h(n)$ is the impulse response of the weighted synthesis filter.

The codebook search procedure attempts to find the values of the adaptive codebook gain G_1 and delay α as well as the fixed codebook index k and gain G_2, which minimise the mean square error E_w taken over the sub-frame length N. This error can be written as

$$E_w = \frac{1}{N}\left(\sum_{n=0}^{N-1} x^2(n) - T_{\alpha k}\right) \qquad (6)$$

where [13]

$$\begin{aligned} T_{\alpha k} &= 2\left(G_1 C_\alpha + G_2 C_k - G_1 G_2 Y_{\alpha k}\right) \qquad (7) \\ &\quad -G_1^2 \xi_\alpha - G_2^2 \xi_k \end{aligned}$$

is the term to be maximised by the codebook search and $x(n) = s_w(n) - \hat{s}_o(n)$ is the target signal for the codebook search. Here

$$\xi_\alpha = \sum_{n=0}^{N-1} [u(n-\alpha) * h(n)]^2 \qquad (8)$$

is the energy of the filtered adaptive codebook signal and

$$C_\alpha = \sum_{n=0}^{N-1} x(n)[u(n-\alpha) * h(n)] \qquad (9)$$

is the correlation between the filtered adaptive codebook signal and the codebook target $x(n)$. Similarly,

ξ_k is the energy of the filtered fixed codebook signal $[c_k(n) * h(n)]$, and C_k is the correlation between this and the target signal. Finally,

$$Y_{\alpha k} = \sum_{n=0}^{N-1} [u(n-\alpha) * h(n)][c_k(n) * h(n)] \quad (10)$$

is the correlation between the filtered signals from the two codebooks.

The usual approach pursued in finding the codebook parameters is to initially set $G_2 = 0$ in Equation 7. Then for a given value of α the optimum gain G_1 can be found by setting the partial derivative of $T_{\alpha k}$ with respect to G_1 to zero. Using this we can then find the value of $T_{\alpha k}$ for every value of α, and choose the adaptive codebook delay which maximises $T_{\alpha k}$. The adaptive codebook parameters are then fixed and a similar procedure is used to find the fixed codebook parameters k and G_2.

In this treatise three sub-optimum techniques are proposed (Methods A...C) and compared to the usual sequential approach as well as to an exhaustive joint search of both codebooks. Setting the partial derivatives of $T_{\alpha k}$ with respect to G_1 and G_2 to zero gives a pair of simultaneous equations which can be solved to give the optimum values of the gains for a given pair of codebook indices α and k. These values are

$$G_1 = \frac{C_\alpha \xi_k - C_k Y_{\alpha k}}{\xi_\alpha \xi_k - Y_{\alpha k}^2} \quad (11)$$

and

$$G_2 = \frac{C_k \xi_\alpha - C_\alpha Y_{\alpha k}}{\xi_\alpha \xi_k - Y_{\alpha k}^2}. \quad (12)$$

The full search procedure computes the terms ξ_α ξ_k C_α C_k and $Y_{\alpha k}$ for every pair of codebook indices α, k and uses these to calculate the gains G_1 and G_2. These gains can then be substituted into Equation 7 to give $T_{\alpha k}$ which the encoder has to maximise by the proper choice of α and k. Most of the complexity of the full search arises from the need to find the cross-correlation term $Y_{\alpha k}$ for each pair of codebook indices. The use of an algebraic fixed codebook structure [2] allows this term, along with ξ_k and C_k, to be found efficiently using a series of four nested loops [6].

The performance of our 4.8kbps ACELP codec, expressed in terms of the segmental signal-to-noise ratio (SEGSNR), for both the usual and the full search procedures is shown in Table 1. Also shown in this table are the performances and relative complexities of various alternative search procedures we simulated. These are

- Method A. In this approach we find α and k with the usual search procedure, and then use Equations 11 and 12 to jointly optimize the values of the codebook gains.

	SEGSNR (dB)	Complexity
Sequential Search	9.7	1
Method A	10.1	1.02
Method B	10.3	1.3
Method C	10.6	1.4
Full Search	10.8	60

Table 1: Performance and Complexity of Various Search Procedures

- Method B. We find the adaptive codebook delay α assuming $G_2 = 0$, and then use only this value of α during the fixed codebook search in which G_1, G_2 and k are all jointly determined. This is similar to an approach suggested in [14] where a very small (32 entries) fixed codebook was used, and a one tap IIR filter was used instead of the adaptive codebook.
- Method C. We find α, k, G_1 and G_2 as in Method B, and then once k is known α is updated by finding G_1, G_2 and $T_{\alpha k}$ for each possible α, and choosing the delay α which maximises $T_{\alpha k}$.

It can be seen from Table 1 that the full joint codebook search offers an improvement of about 1 dB over the sequential codebook search. However it increases the complexity of the codec by a factor of sixty. Method C, which increases the coder complexity by only about 40%, gives almost as good performance as the full search. Finally it can be seen that even Method A, which increases the codec complexity by only 2%, yields a significant performance improvement.

4 Conclusions

In this work we have studied two ways of improving the Analysis-by-Synthesis loop in CELP codecs. Several papers have appeared in the past considering updating the synthesis filter parameters once the excitation signal is known. We found the maximum segmental SNR gain possible through such an update, and proposed a method which obtains almost this full improvement but has a lower complexity than other methods which have been proposed [4, 8]. A significant improvement in the codec's performance can be achieved, similar to the improvement that is obtained with the commonly used techniques of LSF interpolation and bandwidth expansion. Although our LSF update results in a codec that is slightly more complex than one using interpolation and bandwidth expansion, it does not increase the delay of the encoder as interpolation schemes usually do.

Secondly we studied ways of improving the joint adaptive and fixed codebook closed loop searches. Again, initially we found the maximum improvement possible by carrying out a full joint codebook

search. A gain of just over 1dB was achieved over the usual sequential search procedure. We then suggested three sub-optimal search methods. The simplest of these increases the codec's complexity by only 2% but gives almost half a decibel improvement in the codec's segmental SNR. The most complex of the three increases the codec complexity by 40% but gives almost the same improvement in performance as the full joint codebook search.

5 Acknowledgements

The financial support of the SERC, UK (GR/J46845) and that of the Department of Education, Northern Ireland is gratefully acknowledged.

References

[1] Bishnu S. Atal and Joel R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," *Proc. ICASSP*, pp. 614–617, 1982.

[2] C. Laflamme et al., "On Reducing the Complexity of Codebook Search in CELP through the Use of Algebraic Codes," *Proc. ICASSP*, pp. 177–180, 1990.

[3] Sharad Singhal and Bishnu S. Atal, "Optimizing LPC Filter Parameters for Multi-Pulse Excitation," *Proc. ICASSP*, pp. 781–784, 1983.

[4] F.F. Tzeng, "Near-Optimum Linear Predictive Speech Coding," *IEEE Global Telecommunications Conference*, pp. 508.1.1–508.1.5, 1990.

[5] Mahesan Niranjan, "CELP Coding with Adaptive Output-Error Model Identification," *Proc. ICASSP*, pp. 225–228, 1990.

[6] Raymond Steele, *Mobile Radio Communications.* Pentech Press, 1992.

[7] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery, *Numerical Recipes in C.* Cambridge University Press, 1992.

[8] M. Fratti, G.A. Miani and G. Riccardi, "On The Effectiveness of Parameter Reoptimization in Multipulse Based Coders," *Proc. ICASSP*, vol. 1, pp. 73–76, 1992.

[9] Gene H. Golub and Charles F. Van Loan, "An Analysis of the Total Least Squares Problem," *SIAM Journal of Numerical Analysis*, vol. 17, no. 6, pp. 883–890, 1980.

[10] Ronald D. Degroat and Eric M. Dowling, "The Data Least Squares Problem and Channel Equalization," *IEEE Transactions on Signal Processing*, pp. 407–411, 1993.

[11] MD. Anisur Rahham and Kai-Bor Yu, "Total Least Squares Approach for Frequency Estimation Using Linear Prediction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 1440–1454, 1987.

[12] Yoh'Ichi Tohkura, Fumitada Itakura and Shin'Ichiro Hashimoto, "Spectral Smoothing Technique in PARCOR Speech Analysis-Synthesis," *IEEE Trans. on Acoustics, Speech and Signal Processing*, pp. 587–596, 1978.

[13] Jason P. Woodard, "Digital Speech Coding." Mini-Thesis, Department of Electronics and Computer Science, University of Southampton, June 1994.

[14] P. Kabal, J.L. Moncet and C.C. Chu, "Synthesis Filter Optimization and Coding: Applications to CELP," *Proc. ICASSP*, vol. 1, pp. 147–150, 1988.