*Proc. European Signal Processing Conf.* (Edinburgh, Scotland), pp. 908-911, Sept. 1994

# An Enhanced Adaptive Codebook for a CELP Coder

Yasheng QIAN†, Gebrael CHAHINE and Peter KABAL

Dept. of Electrical Eng., McGill University, Montreal, Quebec H3A 2A7
*Tel/Fax: 514 398 7130/ 398 4470, E-Mail:kabal@tsp.ee.mcgill.ca*
†Dept. of Electronic Eng., Tsinghua University, Beijing, China, 100084
*Tel/Fax: 861 259 4690/ 256 4176, E-Mail:qian@tsp.ee.mcgill.ca*

**Abstract.** An adaptive codebook for a one-tap pitch filter has been used for determining the pitch filter using an analysis-by-synthesis procedure in CELP coders. In this paper, we first present the formulations for designing an enhanced adaptive codebook for a pseudo-three-tap pitch synthesis filter, which gives a better performance than a conventional one-tap pitch filter. Then, we focus on the stability analysis of the pseudo-three-tap pitch filters. We propose a sufficient test condition with a relaxed stability, which gives a better performance than a strict stability check. We have employed the enhanced adaptive codebook based on a pseudo-three-tap pitch filter with fractional pitch lags for a 4.8 kb/s CELP speech coder. Both objective and subjective quality have been improved with the enhanced adaptive codebook.

## 1. Introduction

An adaptive codebook for representing a one-tap pitch filter has been successfully used for determining the pitch filter using an analysis-by-synthesis procedure in CELP coders [1], [2]. This analysis-by-synthesis approach provides a better reconstructed speech quality than if the pitch synthesis filter parameters are determined from the input speech. The pitch filter in a low-bit rate speech coder has a strong impact on the performance. Earlier we have reported that a pseudo-three-tap pitch prediction filter is an efficient way to characterize the periodicity in a speech signal [3]. It gives a higher prediction gain and a more appropriate frequency response than a conventional one-tap pitch filter. In contrast to this pitch prediction filter used for speech analysis, a pitch synthesis filter, which is the inverse filter of the pitch prediction filter, is used in speech coders. The adaptive codebook for a pseudo-three-tap pitch synthesis filter is referred to as an enhanced adaptive codebook.

Stability was studied as an important issue for pitch synthesis filters determined by analyzing the input speech [4]. An unstable pitch filter enhances the coding noise. For the analysis-by-synthesis configurations, it has been argued that stability is not important since the choice of filter parameters is based on the reconstructed speech which includes the effect of noise enhancement. Our experimental results, however, show that even in analysis-by-synthesis configurations, stability remains an issue that must be considered.

In our experimental work with unquantized pitch gains, we have seen the pitch coefficients rise to values as high as 800 in transition regions (unvoiced to voiced). In one utterance we saw the average SNR for a CELP coder using an adaptive codebook with unquantized pitch coefficients drop from 7.80 dB for a one-tap filter to 3.89 dB for a three-tap filter. The resulting speech contained annoying pops, clicks and a more dominant background noise. Because we have imposed constraints on the prediction coefficients of the pseudo-three-tap pitch filter, the stability conditions and stabilization procedure can be simplified.

We first describe the enhanced adaptive codebook for a pseudo-three-tap pitch synthesis filter. Then, we focus on the stability analysis for the pitch filter. We present a stabilization procedure with a relaxed stability check, which is better than a strict stability check. Finally, performances of the enhanced adaptive codebook in a 4.8 kb/s CELP coder are given.

This research was supported by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Government of Canada.

## 2. An enhanced adaptive codebook

We employ an enhanced adaptive codebook to determine the pitch lag and prediction coefficients of the pseudo-three-tap pitch synthesis filter in a closed-loop search procedure, as shown in **Figure 1.**

A pseudo-three-tap pitch synthesis filter is a three-tap pitch filter, which has certain constraints on the pitch coefficients, as shown in **Figure 2.** Let the three non-zero coefficients of the pitch filter be $\beta_1, \beta_2$ and $\beta_3$. We can restrict this filter with a symmetrical set of coefficients, by assigning

$$\beta_1 = \beta_2 = \alpha\beta, \qquad \beta_2 = \beta. \tag{1}$$

Both $\beta$ and $\alpha$ are optimized for best performance. This filter has two degrees of freedom. We can further restrict the pseudo-three-tap filter to one degree of freedom by fixing the value of $\alpha$.

The notation for pseudo-three-tap pitch filters $n$T$m$DF, means $n$-taps, $m$ degrees of freedom. Thus, a pseudo-three-tap pitch filter with one degree of freedom is denoted as 3T1DF. Conventional one-tap and three-tap pitch filters are denoted as 1T1DF and 3T3DF, respectively.

The enhanced adaptive codebook corresponds to the set of the pseudo-three-tap pitch filter outputs. Normally, the

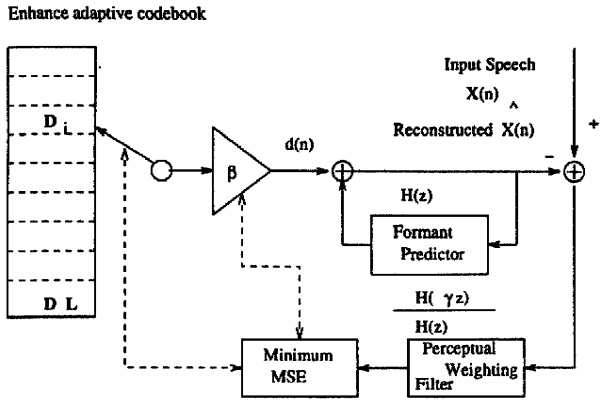


**Figure 1** Close-loop search model with an enhanced adaptive codebook

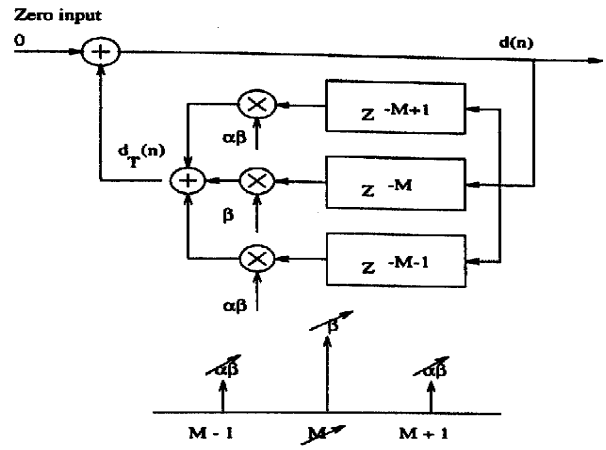


**Figure 2** A pseudo-three-tap pitch synthesis filter

pitch contribution is first chosen with zero codebook excitation. If the subframe length of $N$ is shorter than the delay of $M-1$, the output of the pitch filter, $d(n)$ is obtained by

$$d(n) = \beta\alpha d(n-M-1) + \beta d(n-M) + \beta\alpha d(n-M+1); \quad 1 \le n \le N. \tag{2}$$

The i-th codeword vector of $N$ elements $D_i$ is equal to the $d(n)$ corresponding the given $M$, $\beta$ and $\alpha$. If the subframe length of $N$ is longer than the $M-1$. The other elements, $n > M$, of the codeword repeat the first $M$ samples to give the full-length codeword.

$$\begin{aligned} d(n) &= d(n+M), \quad if \ M \le N \le 2M; \\ d(n) &= d(n+2M), \quad if \ 2M \le N \le 3M. \end{aligned} \tag{3}$$

The delay or the pitch lag, $M$, can be a non-integer for a more accurate representation of the fractional pitch lag. The codeword is, then, obtained by intepolating the available codewords which have integer delays.

The optimal codeword of the enhanced adaptive codebook is determined by minimizing the perceptual weighted mean square errors (MSE) during the closed-loop search. The error between the input speech and reconstructed speech is

$$e(n) = x(n) - \sum_{k=0}^{n} d(k)h(n-k) \tag{4}$$

where h(n) is the impulse response of the formant filter; $d(n)$ corresponds to the codeword, $D_i$. The perceptual weighted error $e_\gamma(n)$ is the convolution of the error $e(n)$ and the impulse response of the perceptual weighted filter $h_\gamma(n)$. There are 128 codewords with integer pitch lags and 128 codewords with non-integer lags, as defined in the Federal Standard 1016 [2].

## 3. Stability

We first explore the stability of the pitch synthesis filter, as determined by an analysis-by-synthesis search procedure. The output of the pitch filter depends on the output of the previous subframe. We can decompose the output into two components: one excited by an ideal prediction residual (at analysis stage), and a quantization noise output.

For the prediction residual, stability is not a problem because of pole/zero cancellation in the analysis and synthesis phase. However, the quantization noise passes through only the unstable synthesis filter. We model the quantization noise to be an additive noise (possibly correlated with the prediction residual). An unstable filter can result in a large boost in the output noise energy. Therefore, the augmented noise can result in pitch filter parameter errors during searching of the adaptive codebook. Furthermore, the noise can be amplified for the consecutive subframes, because the adaptive codebook is updated with the accumulated noise of an unstable pitch filter.

The average SNR for the conventional (unquantized pitch coefficient) 3T3DF pitch filter (one testing sentence), using the closed-loop search, drops down to 3.89 dB, comparing to 9.0 dB for 1T1DF, and produces annoying pops, clicks and dominant background noise. The waveform of the reconstructed speech with an adaptive codebook for a 3T3DF is shown in **Figure 3**. Comparing to the original speech waveform in **Figure 4**, we find that the unstable 3T3DF filter severely impairs the reconstructed speech. In order to alleviate the unstable problem, two stability sufficient test formulas and stabilization techniques have been proposed to efficiently reduce the effect of an unstable pitch filter in [4]. The simple sufficient stability conditions are :

$$\begin{aligned} |\beta| &< 1, \qquad 1T1DF \\ |\beta_1| + |\beta_2| + |\beta_3| &< 1. \qquad 3T3DF \end{aligned} \tag{5}$$

Let $a = \beta_1 + \beta_3$ and $b = \beta_1 - \beta_3$. The tight sufficient stability conditions for a 3T3DF pitch filter are [4]:

$$\begin{aligned} &(1). \ \text{if} \quad |a| \ge |b|, \qquad |\beta_1| + |\beta_2| + |\beta_3| < 1. \\ &(2). \ \text{if} \quad |a| < |b|, \quad \text{and} \quad |\beta_2| + |a| < 1; \\ &\qquad b^2 \le a, \ \text{or} \ b^2\beta_2^2 - (1-b^2)(b^2-a^2) < 0. \end{aligned} \tag{6}$$

The tight sufficient conditions degrade into the simple sufficient conditions ( 5) for both 3T1DF and 3T2DF, since

$b = 0$ and $|a| > 0$. The 3T1DF pitch filter gives a better stability performance than a conventional 3T3DF filter, since we constrain the side prediction coefficients $\beta_1 = \beta_3$ to be a small proportion of the center coefficient $\beta_2$. Let $\alpha = \beta_1/\beta_2$. Therefore, the 3T1DF filter meets the sufficient condition for the simplest stability test in (5)

$$|\beta_2| < \frac{1}{1+2|\alpha|}$$

For a 3T2DF pitch filter with $\beta_1 = \beta_3 = \gamma$, the simplest sufficient condition is

$$2|\gamma| + |\beta_2| < 1$$

A simple stabilization method of the scaled-down coefficients is utilized to stabilize the pitch synthesis filter, if unstable. We scale down the pitch coefficients by multiplying a factor $c$,

$$c = \frac{T_h}{(|\beta_1|+|\beta_2|+|\beta_3|)}, \quad \text{if} \quad (|\beta_1| + |\beta_2| + |\beta_3|) > T_h.$$

The threshold $T_h$ is an experimentally determined threshold. With $T_h = \infty$ no stabilization method is used. With $T_h = 1$, a strict stability condition is imposed. **Figure 5** is the waveform for a simple stabilized 3T3DF$_{b1.0}$. Comparing to the impaired waveform of the unstable 3T3DF filter (**Figure 3**), we find that the improvement with the stabilization method is very good.

## 4. Performance of the enhanced adaptive codebook

The enhanced adaptive codebook for pseudo-three-tap pitch filters, 3T1DF and 3T2DF pitch filters with unquantized coefficients were incorporated into a FS1016 CELP coder. The block diagram of the improved CELP speech coder is depicted in **Figure 6**. The conventional adaptive codebook is replaced by the enhanced adaptive codebook. Other blocks in the speech coder are the same as the FS1016 standard. We employ three performance measures: the average SNR–signal-to-noise ratio, the SEGSNR–segmental signal-to-noise ratio (average of log SNR's evaluated for 16 ms segments and the SFG–synthesis-filter-gain. We define the SFG as the ratio of the energy of the original speech signal and the energy of error between the original speech and the reconstructed speech signal using only the adaptive codebook excitation for the formant synthesis filter. A high value of the SFG indicates that the pitch filter is contributing a large part of the reconstructed signal, while the stochastic codebook is contributing a relatively small part.

Table 1 shows these performance measures for two male and two female test sentences. For comparison, a conventional one-tap filter (1T1DF) and a three-tap filter (3T3DF) are also included. The coefficients are unquantized and the pitch lags are integers, but stabilization as described above is applied. The adaptive codebook for the 3T1DF$_{b2.0}$ obtain a significant increment of SNR gain of 1.16 dB over 1T1DF$_{b1.0}$. The stability threshold $T_h$ is set to be 1.0, 1.10, 1.15 and 1.20 for comparisons. The threshold $T_h$ is denoted in the subscript of the type of the pitch filter. For example, The 3T1DF$_{b1.15}$ employs thresholds of 1.15. The 3T1DF$_{b\infty}$ uses the $T_h = \infty$. It means that the pitch filter is not stabilized.

The results show that the stabilization actually improves the performance. Moreover, a relaxed stability constraint is better than a strict stability constraint. The reason is that the increasing pitch pulse amplitudes are better to model the fast growing voicing onset. The SNR for 3T1DF$_{b2.0}$ is higher than the 1T1DF$_{b1.0}$ by 1.16 dB. The SNR difference between the 3T1DF$_{b2.0}$ and the 3T3DF$_{b2.0}$ is small (0.32 dB).

We have also applied quantization to the 3T1DF pitch filter coefficients. The quantization table is defined in the FS1016 CELP coder specification. Notice that the stabilization is in effect present, since the largest quantized value for $|\beta_2|$ is 1.991. Therefore, the maximum sum of $|\beta_2|(1+2|\alpha|)$ = 2.53, because we select $\alpha = 0.135$. With quantization, the SNR for the 3T1DF$_{b2.0}$ configuration drops by only 0.13 dB.

Finally, we have evaluated the SNR and SEGSNR for the 3T1DF pitch filter with fractional pitch lags [5], [6] and pitch quantizer (FS1016 CELP coder). The results show that the SNR and SEGSNR increase by 0.44 dB and 0.05 dB, respectively, over those of the integer pitch filter. The SNR and SEGSNR are higher than standard FS1016 coder by 0.45 dB and 0.1 dB. An informal listening test show that the improved CELP coder with 3T1DF pitch filter is better than the original FS1016 CELP coder.

## 5. Conclusions

The enhanced adaptive codebook for pseudo-three-tap pitch synthesis filters can be incorperated in a CELP coder to improve the speech quality. A scaled-down pitch coefficients technique with a relaxed sufficient constraints to obtain a weakly unstable pitch synthesis filter can track fast changing segments during a unvoicing to voicing onset. The performance of the improved 4.8 kb/s CELP coder with the pseudo-three-tap pitch filter is better than the FS1016 coder with a one-tap pitch filter.

## References


[1] **W.B. Kleijn, D.J. Krasinski, and R.H. Ketchum (1988)** Improved speech quality and efficient vector quantization in selp, *Proc. Int. Conf. ICASSP-88, vol I,155–158*, New York, April 1988.

[2] **J.P. Campbell, T.E. Tremain, and V.C. Welch (1990)**, The proposed federal standard 1016 4800 bps voice coder: Celp, *Speech Technology*, **Apr./May**, 58–64.

[3] **Y. Qian and P. Kabal (1993)**, Pseudo-three-tap pitch prediction filters, *Proc. Int. Conf. Acoust., ICASSP-93, vol II, 523-526*, Minneapolis, April 1993.

[4] **R. Ramachandran and P. Kabal (1987)**, Stability and performance analysis of pitch filters in speech coders, *IEEE Trans. on ASSP*, **35**, 937–946.

[5] **P. Kroon and B. S. Atal (1991)**, Pitch predictors with high temporal resolution, *IEEE Trans. Signal Processing*, **39**, 733-735.

[6] **R. Crochiere and L. Rabiner (1983)**: Multirate Digital Signal Processing, *Prentice-Hall International*

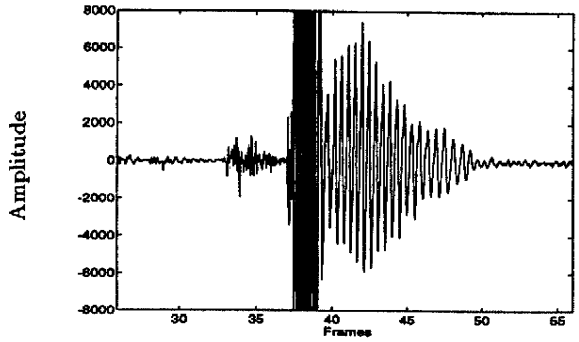


**Figure 3** Reconstructed waveforms with an unstable 3T3DF pitch synthesis filter

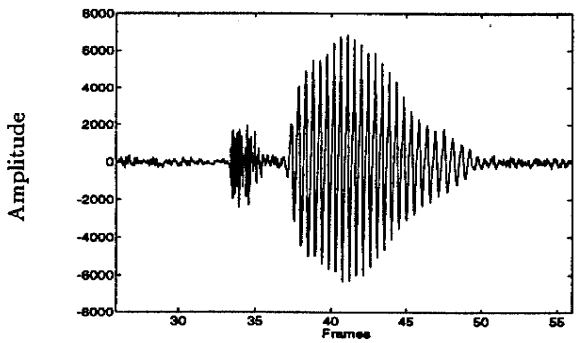


**Figure 4** Original speech waveforms

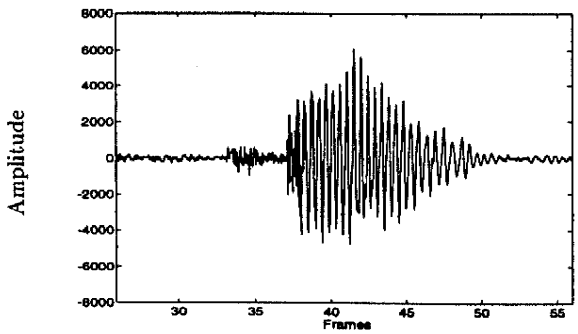


**Figure 5** Reconstructed waveforms with a stabilized $3T3DF_{b1.0}$ pitch filter

| Type | SNR (dB) | SEGSNR (dB) | SFG (dB) |
|---|---|---|---|
| $1T1DF_{b\infty}$ | 7.80 | 7.77 | 5.52 |
| $1T1DF_{b1.0}$ | 7.13 | 7.74 | 5.33 |
| $1T1DF_{b1.1}$ | 7.85 | 7.80 | 5.29 |
| $1T1DF_{b1.15}$ | 7.81 | 7.73 | 5.27 |
| $1T1DF_{b2.0}$ | 7.99 | 7.88 | 5.27 |
| $3T1DF_{b\infty}$ | 6.77 | 7.89 | 5.66 |
| $3T1DF_{b1.0}$ | 7.72 | 7.88 | 5.17 |
| $3T1DF_{b1.10}$ | 8.11 | 7.97 | 5.40 |
| $3T1DF_{b1.15}$ | 8.26 | 8.02 | 5.42 |
| $3T1DF_{b2.0}$ | 8.29 | 8.00 | 5.59 |
| $3T2DF_{b\infty}$ | 4.60 | 8.03 | 5.78 |
| $3T2DF_{b1.0}$ | 6.89 | 7.19 | 4.85 |
| $3T2DF_{b1.1}$ | 7.28 | 7.32 | 5.09 |
| $3T2DF_{b1.15}$ | 7.43 | 7.64 | 5.36 |
| $3T2DF_{b2.0}$ | 8.30 | 8.18 | 5.68 |
| $3T3DF_{b\infty}$ | 3.89 | 8.27 | 5.98 |
| $3T3DF_{b1.0}$ | 7.37 | 7.58 | 4.75 |
| $3T3DF_{b1.15}$ | 7.78 | 7.94 | 5.65 |
| $3T3DF_{b2.0}$ | 8.61 | 8.32 | 5.91 |

Table 1 SNR (dB) comparisons for different pitch synthesis filters in a CELP speech coder

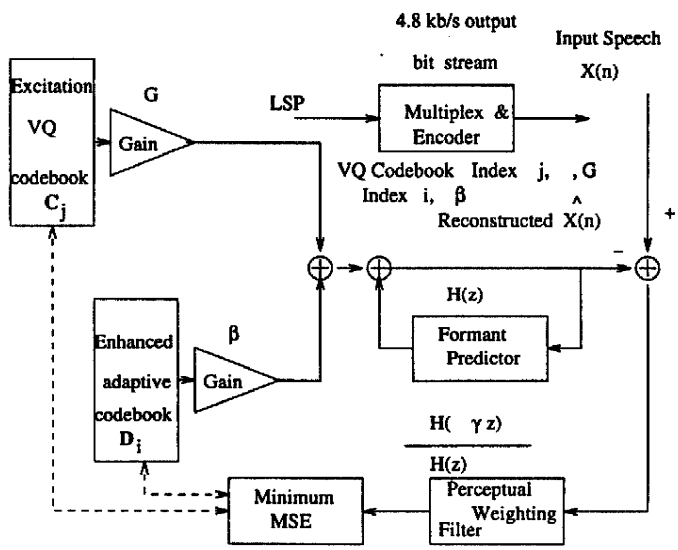


**Figure 6** Block diagram of the improved CELP speech coder