Proc. Int. Conf. Commun. Technol. (Shanghai, China), pp. 24.01.1-24.01.4, June 1994

Speech Coding at 4.8 kb/s with an Improved Pitch Filter

QIAN Yasheng [1] and Peter KABAL [2]

[1] Dept. of Electronic Eng.
Tsinghua University
Beijing, China

[2] Dept. of Electrical Eng.
McGill University
Montreal, Quebec H3A 2A7

Abstract

The reconstructed speech quality in a low bit-rate CELP coder is very dependent on the performance of the pitch filter. In this paper, we present an improved pitch filter, a fractional pseudo-three-tap pitch synthesis filter, which performs better than a conventional one-tap pitch filter. We discuss the frequency response of the improved pitch filter. We explore stability issues for three-tap pitch filters in a CELP coder. We have incorperated a fractional pseudo-three-tap pitch filter into a 4.8 kb/s CELP speech coder. Both objective and subjective quality have been improved with the fraction pseudo-three-tap pitch filter.

1. Introduction

We have reported that a pseudo-three-tap pitch prediction filter gives a higher prediction gain and a more appropriate frequency response than a conventional one-tap pitch filter [1]. 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 CELP coders.

A fractional pitch filter with high temporal resolution can improve the performance of CELP coders [2]. The pseudo-three-tap pitch filter can also enhance its performance using fractional pitch lags. An analysis-by-synthesis procedure is used to determine the fractional pitch filter parameters.

The frequency response of a one-tap pitch synthesis filter with integer or non-integer lags shows a constant envelope constraining the pitch peaks. Since the pitch prediction coefficients are symmetrical in the pseudo-three-tap pitch filter, the frequency response can be more appropriate than for one-tap or general three-tap pitch filters.

Stability was studied as an important issue for pitch synthesis filters determined by analyzing the input speech in [3]. An unstable pitch filter enhances the coding noise. For the analysis-by-synthesis procedure, the choice of filter parameters is based on the reconstructed speech which includes the effect of noise enhancement. Our experimental results, however, show that stability remains an issue that must be considered.

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

2. A fractional pseudo-three-tap pitch synthesis filter

A fractional pseudo-three-tap pitch synthesis filter is a fractional three-tap pitch filter, which has certain constraints on the pitch coefficients, as shown in Fig. 1. Let the three non-zero coefficients of the pitch filter be β_{-1}, β_0 and β_{+1}. We can restrict this filter to a symmetrical set of coefficients, by assigning

$$\beta_{-1} = \beta_{+1} = \alpha\beta, \qquad \beta_0 = \beta. \tag{1}$$

Both β and α 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 α.

The notation adopted for pseudo-three-tap pitch filters is nTmDF, meaning 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.

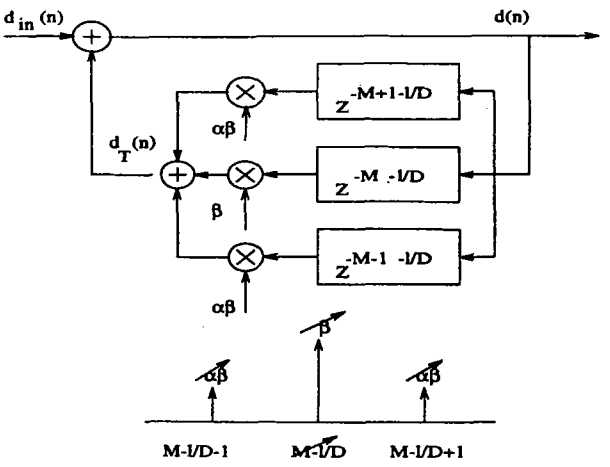

Fig. 1 A fractional pseudo-three-tap pitch synthesis filter

The non-integer pitch lag can be expressed as an integer number of samples plus a fraction of a sampling interval. Let the pitch resolution be $1/D$. The fractional part of the pitch lag can be expressed as l/D, where

$l = 0, 1, \ldots, D-1$, $(1 > l/D \geq 0)$. The pseudo-three-tap filter then acts on the interpolated samples, denoted by $d^{(l)}(n-(M-1)), d^{(l)}(n-M), d^{(l)}(n-(M+1))$.

A polyphase filter structure [4] can be used to obtain the interpolated samples $d^{(l)}(n)$. For each fractional phase l, the impulse response $p^{(l)}(n)$ of the polyphase filter is obtained by sub-sampling an appropriate interpolating filter $h_i(n)$. In our case, we use an interpolation filter which is a Hamming-windowed low-pass filter,

$$p^{(l)}(n) = w_h(n - l/D)\frac{\sin(\pi(n-l/D)}{\pi(n-l/D)}, \tag{2}$$

where $w_h(n)$ is a Hamming window (centered at zero).

The interpolated sample at $n + l/D$ is given by,

$$d^{(l)}(n) = \sum_{k=0}^{q-1} p_l(k-I)d(n-k), \tag{3}$$

where $q = 2I$ is the number of the coefficients of the polyphase filter[1]; I is the delay of the causal interpolation filter at the original sampling rate. The resulting value which corresponds to the output of the fractional pseudo-three-tap pitch synthesis filter for the pitch lag of $M - l/D$ can be written as

$$d(n) = d_{in}(n-I) + \sum_{i=-1}^{1}\sum_{k=0}^{q-1}\beta_i\, p_l(k)\, d(n-(M+i)-k). \tag{4}$$

where d_{in} is the input to the pitch filter, i.e., the code vector multiplied by the gain factor G

We employ a closed-loop sequential search procedure to determine the fractional pitch lag and prediction coefficients of the fractional pseudo-three-tap pitch synthesis filter, as shown in Fig. 2.

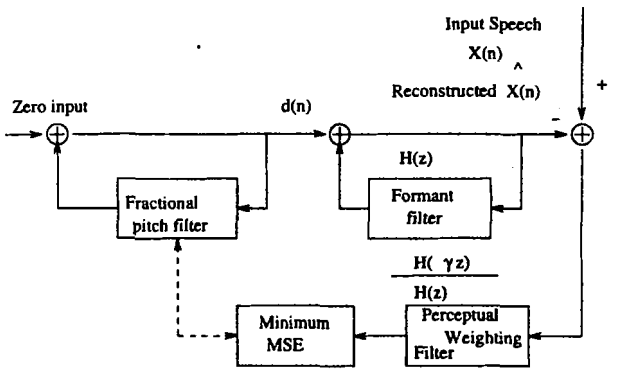

Fig. 2 Closed-loop sequential search for a fraction pitch filter

For the 3T1DF case, we define

$$d_\alpha^{M-l/D}(n) = \alpha d(n-M-l/D-1) + d(n-M-l/D) + \alpha d(n-M-l/D+1) \tag{5}$$

[1] q should be $2I+1$ for $l = 0$. However, for this case, there is only one non-zero coefficient for the case.

For a zero codebook contribution, $d_{in}(n-I) = 0$,

$$d(n) = \beta d_\alpha^{M-l/D}(n) \tag{6}$$

The error signal between the input speech and reconstructed speech is

$$\begin{aligned} e(n) &= x_w(n) - \sum_{j=0}^{n} d_w(j)h(n-j) \\ &= x_w(n) - \beta\sum_{j=0}^{n} d_{w\alpha}^{M-l/D}(j)h(n-j). \end{aligned} \tag{7}$$

where $h(n)$ is the impulse response of the format filter; $x_w(n)$ is the windowed input; $d_w(n)$ is $d(n)$ multiplied by the data window. We multiply an error window $w_e(n)$ to obtain a windowed error signal $e_w(n)$. The resulting summed squared error is

$$\varepsilon = \sum_{n=-\infty}^{\infty} e_w^2(n). \tag{8}$$

In our block-based analysis, we use a covariance analysis with $w_d(n) = 1$ for all n and a rectangular error window $w_e(n) = 1$ for $0 \leq n \leq L-1$. The $M - l/D$ is chosen as that which is optimal for a one-tap pitch filter. The perceptual weighted error $e_{w\gamma}(n)$ is the convolution of the error $e_w(n)$ and the impulse response of the perceptual weighted filter $h_\gamma(n)$.

The optimal pitch filter parameters (β and $M - l/D$) can be obtained by minimizing ε. Setting partial derivatives of ε to zero, we obtain the analytical optimum β_{opt}.

$$\beta_{opt} = \frac{\sum_{n=-\infty}^{\infty} x_w(n)\sum_{j=0}^{n} d_{w\alpha}^{M-l/D}(j)h(n-j)}{[\sum_{j=0}^{n} d_{w\alpha}^{M-l/D}(j)h(n-j)]^2} \tag{9}$$

After choosing the pitch filter parameters with zero input, the optimal excitation is determined by minimizing the perceptually weighted mean square errors (MSE) during a closed-loop search (7).

3. Frequency response

The reconstructed speech spectrum depends on the frequency response of the pitch synthesis filter and format filter. We will discuss and compare the frequency response of fractional pseudo-three-tap synthesis filters 3T1DF with conventional 1T1DF and 3T3DF filters.

The frequency response of a 3T3DF pitch filter is expressed as

$$H(e^{j\omega}) = \frac{1}{1 - \beta_{-1}e^{j\omega(M-1)} - \beta_0 e^{j\omega M} - \beta_{+1}e^{j\omega(M+1)}} \tag{10}$$

Then, the amplitude of frequency response of a 3T3DF pitch filter can be written as

$$\begin{aligned} |H(e^{j\omega})| &= \frac{1}{\sqrt{R(w)^2 + Im(w)^2}} \\ R(w) &= \cos(\omega M) - \beta_0 - (\beta_{-1} + \beta_{+1})\cos(\omega) \\ Im(w) &= (\beta_{+1} - \beta_{-1})\sin(\omega) + \sin(\omega M) \end{aligned} \tag{11}$$

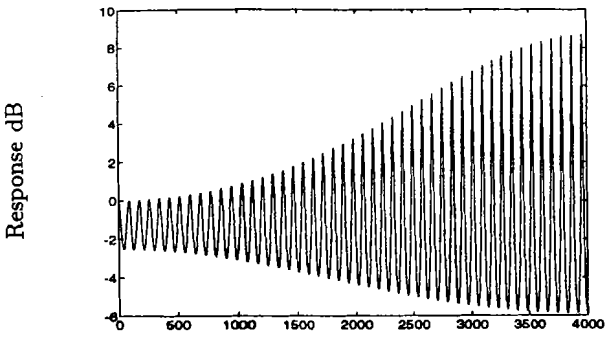

Fig. 3 Frequency responses of a three-tap pitch filter with coefficients (–0.14,0.41,–0.14)

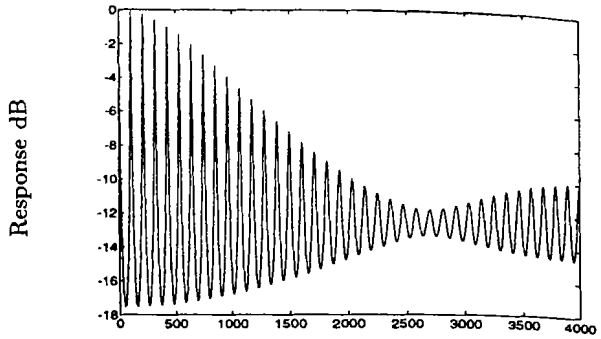

Fig. 4 Frequency responses of a three-tap pitch filter with coefficients (0.31,0.25,0.20)

Since pitch period M is in the range of 20 – 147, we consider $M >> 1$. Terms of $\cos(\omega M)$ and $\sin(\omega M)$ produce the quasi-harmonics structure in the frequency response. The envelope of the frequency response mainly depends on the terms of $(\beta_{-1} + \beta_{+1})\cos(\omega)$ and $(\beta_{+1} - \beta_{-1})\sin(\omega)$. The term $\cos(\omega)$ is a monotonic decreasing function from 1 to -1, corresponding to $\omega = (0, \pi)$. The term $(\beta_{+1} - \beta_{-1})\sin(\omega)$ reaches maximum of $(\beta_{+1} - \beta_{-1})$ at $\omega = \pi/2$. For a given pitch period M, the envelope depends on the values of $\beta_{-1}, \beta_0, \beta_{+1}$. There are four possible envelopes:

1. A decreasing monotonic shape, if $\beta_0 > (\beta_{-1} + \beta_{+1}) > 0$.;
2. An increasing monotonic envelope, Fig. 3, if $(\beta_{-1} + \beta_{+1}) < 0$ and $|\beta_{-1} + \beta_{+1}| \approx \beta_0$;
3. Two resonances, Fig. 4, if $(\beta_{-1} + \beta_{+1}) > \beta_0 > 0$. and $\beta_{+1} > \beta_{-1}$; The term $(\beta_{+1} - \beta_{-1})\sin(\omega)$ makes an important contribution in the middle region. Since this term vanishes at the $\omega = 0, \pi$, there is a valley in the middle region.
4. A resonance in the middle, if β_{-1} and β_{+1} have different signs;

For the fractional 3T1DF case, $|H(e^{j\omega})|$ becomes

$$|H(e^{j\omega})|_{\text{3T1DF}} = \frac{1}{\sqrt{R_{\text{3T1DF}}(w)^2 + Im_{\text{3T1DF}}(w)^2}}$$
$$R_{\text{3T1DF}}(w) = \cos(\omega(M - l/D)) - \beta(1 + 2\alpha\cos(\omega)$$
$$Im_{\text{3T1DF}}(w) = \sin(\omega(M - l/D) \quad (12)$$

The amplitude of $|H(e^{j\omega})|$ of 3T1DF has only a decreasing envelope. The frequency response of the fractional 3T1DF filter with $\alpha = 0.250$ is shown in Fig. 5.

Let the $\alpha = 0$ in (12). Then, the $|H(e^{j\omega})|$ becomes the constant envelope of a 1T1DF pitch filter, Fig. 6.

$$|H(e^{j\omega})| = \frac{1}{\sqrt{1 - 2\beta\cos(\omega(M - l/D)) + \beta^2}} \quad (13)$$

Comparing to Figs. 6, we find that Fig. 5 of the 3T1DF pitch filter response is more desirable than that of the 1T1DF.

4. Stabilization

We explore the stability for the pitch synthesis filter, determined by an analysis-by-synthesis search procedure. 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.

Two stability sufficient test formulas and stabilization techniques have been proposed to efficiently reduce the effect of an unstable pitch filter in [3]. 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.

For a 3T1DF pitch filter, the simplest sufficient stability condition is

$$|\beta_0| < \frac{1}{1+2|\alpha|}$$

For a 3T2DF pitch filter with $\beta_{-1} = \beta_{+1} = \gamma$, the sufficient condition is

$$2|\gamma| + |\beta_0| < 1$$

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

$$c = \frac{Th}{(|\beta_{-1}| + |\beta_0| + |\beta_{+1}|)}, \quad \text{if } (|\beta_{-1}| + |\beta_0| + |\beta_{+1}|) > Th.$$

The threshold Th is an experimentally determined threshold.

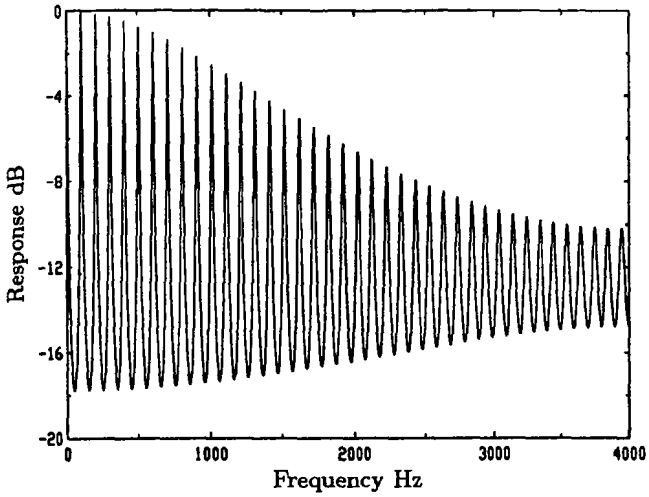

Fig. 5 Frequency response of a 3T1DF pitch synthesis filter with $\alpha = 0.25$, pitch = 78 Hz

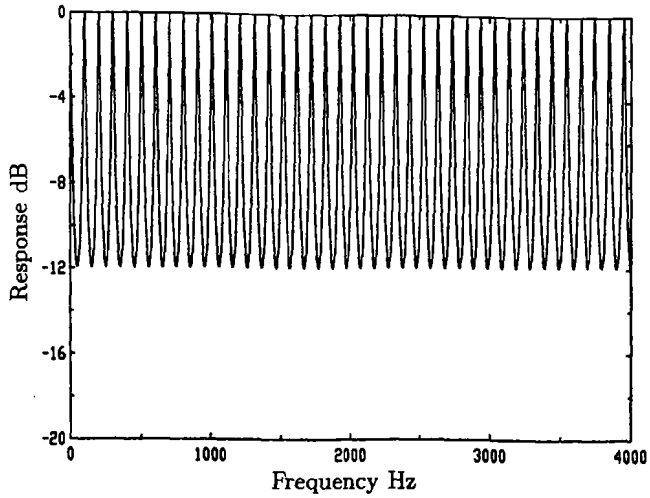

Fig. 6 Frequency response of a one-tap pitch synthesis filter

5. Performance

The fractional pseudo-three-tap pitch filters, 3T1DF and 3T2DF pitch filters are incorporated into a FS1016 4.8 kb/s CELP coder. We employ two performance measures: the average signal-to-noise ratio (SNR) and the segmental signal-to-noise ratio (SEGSNR). They are the average of log SNR's evaluated for 16ms segments. We have tested the performance for two male and two female sentences. The prediction coefficients of the pitch filter are first unquantized and the pitch lags are integers, but stabilization as described above is applied. The stability threshold Th is set to be 1.00, 1.10, 1.15, 2.00 and ∞ for comparisons. The threshold Th is denoted in the subscript of the type of the pitch filter. For example, $3T1DF_{b1.15}$ employs thresholds of 1.15, while $3T1DF_{b\infty}$ uses the $Th = \infty$. This 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 able to model a fast growing voicing onset. The SNR for $3T1DF_{b1.15}$ is higher than the $1T1DF_{b1.15}$ by 1.13 dB. It is higher than $3T3DF_{b1.15}$ by 0.46 dB. The SNR for the $3T3DF_{b2.00}$ has 0.32 dB more than $3T1DF_{b2.00}$.

We have also applied quantization to the 3T1DF pitch filter coefficients. The quantization table is defined in the FS1016 CELP coder specification. Notice that 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 $3T1DF_{b2.00}$ configuration drops by only 0.13 dB.

Finally, we have evaluated the SNR and SEGSNR for the 3T1DF pitch filter with fractional pitch lags 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. An informal listening test show that the improved CELP coder with 3T1DF pitch filter is better than the original FS1016 CELP coder.

6. Conclusions

The fractional pseudo-three-tap pitch synthesis filters can be incorporated into 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 fractional pseudo-three-tap pitch filter is better than the FS1016 coder with a one-tap pitch filter.

References

[1] Y. Qian and P. Kabal, "Pseudo-three-tap pitch prediction filters," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, (Minneapolis, MN), pp. II–523–II–526, April 27–30 1993.

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

[3] R. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans. on Acoustics, Speech, Signal Processing*, vol. 35, pp. 937–946, July 1987.

[4] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1983.