Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Minneapolis, MN), pp. II-523-II-526, April 1993

PSEUDO-THREE-TAP PITCH PREDICTION FILTERS

Qian Yasheng [1,2] Peter Kabal [1]

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

[2] Dept. of Electronic Eng.
Tsinghua University
Beijing, China
100084

ABSTRACT

Pitch filters play an important role in high quality medium and low rate speech coders. We propose a pseudo-three-tap pitch filter with one or two degrees of freedom of the prediction coefficients, which gives a higher pitch prediction gain and also a more desirable frequency response than a one-tap pitch prediction filter. First, we describe an analysis model for the pseudo-three-tap pitch filter. Then, we apply the pseudo-three-tap concept together with a fractional pitch lag. The pitch prediction gain and frequency response of the pseudo-three-tap pitch filters are compared to a one-tap and three-tap pitch predictors with an integer or a non-integer pitch lag. The pseudo-three-tap pitch filter with one degree of freedom outperforms a conventional one-tap pitch filter. Even better is a pseudo-three-tap filter which switches between two parameter values.

1. Introduction

Pitch filters combined with a formant filters have been successfully used in medium and low rate high-quality speech coders [1], [2]. More recently, an 8 kb/s Low-Delay CELP speech coder with a cascaded-backward adaptive formant and a three-tap forward-adaptive pitch filter has been presented [3]. A three-tap pitch filter provides better speech quality than a one-tap pitch filter. However, more bits are required to encode the additional two pitch filter coefficients.

The objective of our work is to develop a more efficient way to represent the multi-tap pitch filter. To this end, and to try to draw conclusions applicable to a wide variety of speech coder configurations, we focus our attention on pitch prediction filters. The filter used at the synthesis stage of a speech coder is the inverse of the prediction filter.

We propose a pseudo-three-tap pitch prediction filter, which has three non-zero pitch coefficients with one or two degrees of freedom. These pseudo-three-tap pitch filters can give a higher pitch prediction gain than a one-tap pitch filter.

The frequency response of a one-tap filter shows a constant envelope constraining the pitch peaks (see Fig. 1). The search for pseudo-three-tap pitch filters was motivated by the observation that the spectrum of a conventional three-tap pitch filter often shows a a diminishing envelope with increasing frequency in many voiced segments (see Fig. 2). This corresponds to a large center coefficient and smaller sign side coefficients. Such a frequency response adds more pitch structure at low frequencies than at high frequencies. Note also that if the true pitch corresponds to a half integer lag, the frequency response variations due to an integer lag pitch filter match at low frequencies but become increasingly mismatched to the true pitch peaks until they are 90 degrees out of phase at the half-sampling frequency. A reduced high frequency pitch component will reduce the apparent effect of such mismatched lags.

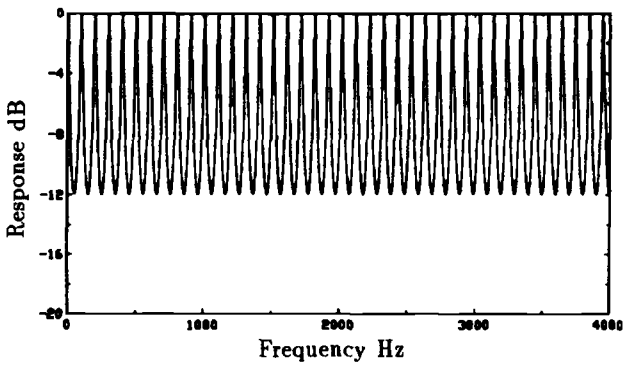

Fig. 1 Frequency response of a one-tap pitch synthesis filter

A conventional view of three-tap pitch filters is that they can interpolate between integer lags. This has led to the development of fractional pitch filters where the interpolation is explicitly carried out [4]. Additional bits are needed to code the resulting higher resolution pitch lags. Note that such one-tap fractional-pitch filters have a constant envelope frequency response.

In this paper, we first describe a general analysis model for the pseudo-three-tap pitch filter. Then, we combine the pseudo-three-tap concept with a fractional pitch lag. Finally we compare the performance of the pseudo-three-tap pitch filter (both integer and fractional lag) with traditional one-tap and three-tap pitch filters.

2. Pseudo-three-tap Pitch Filters

A pseudo-multi-tap pitch filter is an n-tap pitch filter which has fewer than n degrees of freedom. A traditional

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

0-7803-0946-4/93 $3.00

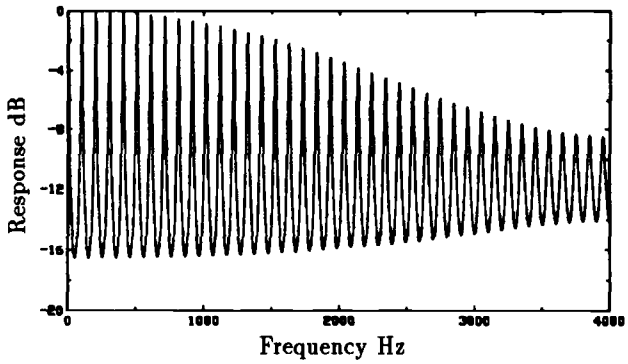

Fig. 2 Frequency response of a three-tap pitch synthesis filter

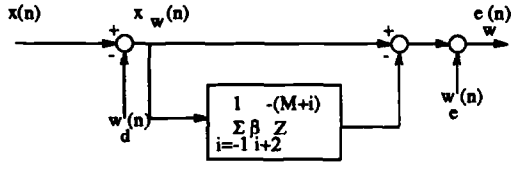

Fig. 3 Analysis model for a pseudo-three-tap predictor

three-tap pitch filter has three degrees of freedom. Here we consider pseudo-three-tap filters with only one or two degrees of the freedom. Let the three non-zero coefficients of the three-tap pitch filter be β_1, β_2, and β_3. We can restrict this filter to two degrees of freedom, while maintaining a symmetrical set of coefficients, by assigning

$$\beta_1 = \beta_3 = \alpha\beta, \qquad \beta_2 = \beta\ . \tag{1}$$

Both β and α are optimized for best performance. We can further restrict the pseudo-three-tap filter to one degree of freedom by fixing the value of α.

The notation for pseudo-multi-tap filters are nTmDF, meaning n taps, m degrees of freedom. Thus, a convention three-tap filter is 3T3DF (β_1, β_2 and β_3 variable). The pseudo-three-tap filters are 3T2DF (α and β variable) and 3T1DF (α fixed, β variable).

An analysis model for calculating the prediction coefficients of the pseudo-three-tap pitch predictor with a transversal implementation is shown in Fig. 3. The input signal $x(n)$ is multiplied by a data window $w_d(n)$ to give $x_w(n)$. The signal $x_w(n)$ is predicted from a set of its previous samples with lags of $M-1$, M, $M+1$. The prediction error is

$$e(n) = x_w(n) - \sum_{i=-1}^{1} \beta_{i+2}\, x_w(n-(M+i)), \tag{2}$$

where M is the pitch lag corresponding to the middle tap. The final step is to multiply the error signal by an error window $w_e(n)$ to obtain a windowed error signal $e_w(n)$. The resulting summed squared prediction error is

$$\varepsilon^2 = \sum_{n=-\infty}^{\infty} e_w^2(n). \tag{3}$$

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 $0 \le n \le L-1$. The lag M is chosen as that which is optimal for a one-tap pitch predictor [2]. The coefficients β_i are computed by minimizing ε^2.

The minimization of ε^2 leads to a set of different linear equations which can be written in matrix form. For the case of 3T2DF, define $\gamma = \alpha\beta$. Then setting partial derivatives of ε^2 to zero,

$$\begin{bmatrix} A & B \\ B & D \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} = \begin{bmatrix} \phi(0,M-1) + \phi(0,M+1) \\ \phi(0,M) \end{bmatrix}, \tag{4}$$

where

$$A = \phi(M-1,M-1) + \phi(M+1,M+1) + 2\phi(M-1,M+1)$$
$$B = \phi(M-1,M) + \phi(M,M+1)$$
$$D = \phi(M+1,M+1)$$

and $\phi(i,j)$ is defined as

$$\phi(i,j) = \sum_{n=-\infty}^{\infty} x_w(n-i)x_w(n-j). \tag{5}$$

Using this formulation, we obtain β_{opt} and γ_{opt},

$$\begin{aligned} \beta_{\text{opt}} &= (AF - BE)/(AD - B^2), \\ \gamma_{\text{opt}} &= (DE - BF)/(AD - B^2). \end{aligned} \tag{6}$$

For the case of 3T1DF,

$$\beta_{\text{opt}} = \frac{\alpha\phi(0,M-1) + \phi(0,M) + \alpha\phi(0,M+1)}{\alpha^2\phi_3 + \phi(M,M) + 2\alpha\phi_2}, \tag{7}$$

where

$$\phi_3 = \phi(M-1,M-1) + 2\phi(M-1,M+1) + \phi(M+1,M+1),$$
$$\phi_2 = \phi(M-1,M) + \phi(M,M+1).$$

3. Fractional Pitch Lags

The use of a fractional pitch lag has proved to be an accurate and efficient means to characterize speech speech periodicity in low bit-rate speech coders. Fractional pitch lags can also be applied to pseudo-three-tap pitch predictors. The non-integer pitch lag can be expressed as an integer number of samples plus a rational 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 $l = 0, 1, \ldots, D-1$. The three-tap filter then acts on the interpolated samples, denoted by $x_w^{(l)}(n-(M-1)), x_w^{(l)}(n-M), x_w^{(l)}(n-(M+1))$.

A polyphase filter structure [5] is used to obtain the interpolated samples. For each phase l, the impulse response $p_l(n)$ is obtained by sub-sampling an appropriate interpolating filter $h(n)$. In our case, we use an interpolated filter which is a Hamming-windowed ideal low-pass filter,

$$p_l(n) = w_h(n-I)\sin\frac{(\pi(n-I-l/D)}{\pi D(n-I-l/D)}, \tag{8}$$

where $w_h(n)$, $-I \le n \le I$, is a Hamming window (centered at zero).

The resulting value which corresponds to the interpolated sample at lag $n + l/D$ is given by,

$$x_w^{(l)}(n) = \sum_{k=0}^{q-1} p_l(k) x_w(n-k), \quad (9)$$

where $q = \lceil (2ID+1)/D \rceil$.

The prediction error signal of the pseudo-three-tap pitch predictor for a (fractional) pitch lag of $M + l/D$ can be written as

$$e(n) = x_w(n) - \sum_{i=-1}^{1} \sum_{k=0}^{q-1} \beta_{i+2}\, p_l(k)\, x_w(n-(M+i)-k). \quad (10)$$

For the fractional pitch case, the optimal pitch predictor parameters can be obtained by minimizing ε^2, as in the previous section, but with the covariance function appropriately modified. The new covariance function with a fractional delay is,

$$\phi^{(l)}(i,j) = \sum_{n=-\infty}^{\infty} \sum_{k=0}^{q-1} p_l(k) x_w(n-i-k) \sum_{k=0}^{q-1} p_l(k) x_w(n-j-k). \quad (11)$$

4. Performance of Pseudo-three-tap Pitch Filters

To compare the pseudo-three-tap pitch filter with conventional one-tap and three-tap pitch filters, the pitch prediction gain is used to measure the performance. The predictor gain is the ratio of the energy at the input to the predictor to that of the prediction error (expressed in dB). In all cases, a forward-adaptive pitch prediction is applied to the residual produced by a forward-adaptive formant prediction filter with 10 taps, updated every 160 samples. The pitch filters themselves are updated every 20 samples. The lag value chosen $M + l/D$ is that which is best for a one-tap pitch filter. With this choice of $M + l/D$, the prediction coefficients are symmetrical about the lag value. Fig 5 shows the average pitch prediction gains for a number of configurations, all with integer pitch lags. The results are shown for a single sentence for male and female speakers. Note that the performance of the 3T1DF configuration depends on the value of α chosen. The results shown in the figure indicate that $\alpha = 0.125$ is good for both male and female speech. The average gains are about 0.2 dB higher than the conventional one-tap pitch filter.

The frequency response of a 3T1DF filter with $\alpha = 0.25$ is shown in Fig. 4. This can be compared to Figs. 1 and 2.

In some speech frames, the 3T1DF pitch filter is better by 1.5–2.0 dB, but in others it can in fact be slightly worse than the 1T1DF configuration. This suggests that it is possible to combine these two configurations, switching to the one which performs the best. Fig. 6 shows the results (3T1DFS, D=1) of switching between $\alpha = 0$ (the 1T1DF case), and another non-zero value. With switching, $\alpha = 0.25$ is preferable to $\alpha = 0.125$. Note that switching between α's costs one bit. This approach can also be considered to coarsely quantize the α parameter of a 3T2DF configuration.

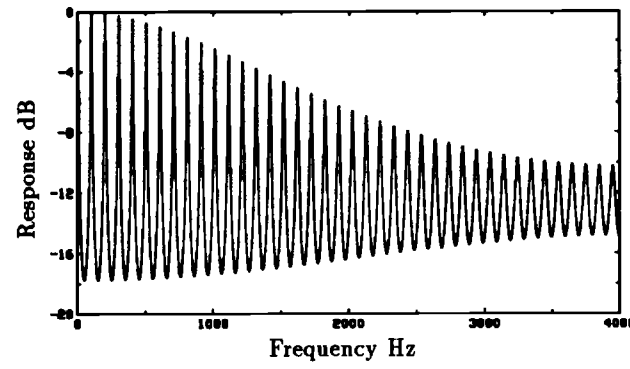

Fig. 4 Frequency response of a 3T1DF pitch synthesis filter with $\alpha = 0.25$

Next we compare the pseudo-three-tap pitch filter with a fractional pitch lag to one-tap and a three-tap conventional pitch filters with a fractional pitch lag. The FIR interpolation filter is selected to have $I = 16$ (16 samples from each side of the desired location are used for the interpolation). A number of different interpolation ratios were used (maximum 16). The pitch prediction gain of a 3T1DF filter as a function of α for various interpolation ratios D is shown in Fig. 5. The pitch prediction gain for 3T1DF with a fractional pitch lag increases with the interpolation ratio as does that for the 1T1DF case. (The 1T1DF case is the same as 3T1DF with $\alpha = 0$.) However, the pitch prediction gain saturates when the interpolation ratio is larger than 8.

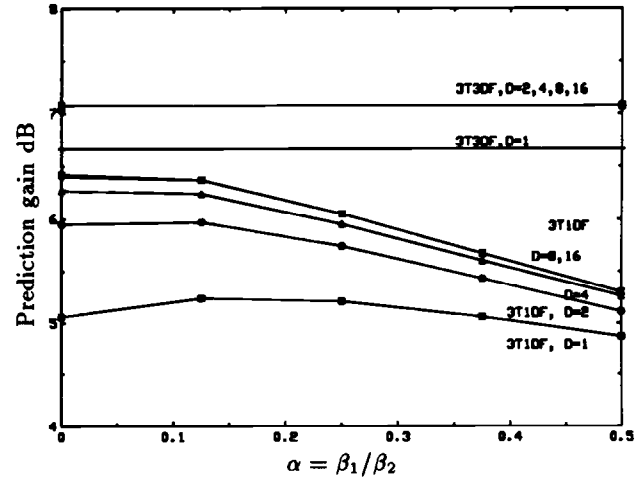

Fig. 5 Pitch prediction gains for pitch filters versus α for different values of D, male speech

We have also evaluated a conventional three-tap pitch filter 3T3DF with a fractional pitch lag. The 3T3DF with an interpolation ratio of $D = 2$ gives an increased prediction gain of 0.41 dB for male speech. The 3T3DF with a higher interpolation ratio $D > 4$ does not provide more pitch prediction gain. This is in contrast with a 1T1DF

filter, where $D = 2$ gives an increase in 0.89 dB. Further smaller increases occur for higher values of D, but with performance levelling off below even the 3T3DF value for $D = 1$. One interpretation of these results is that the 3T3DF filter exploits the redundancy among three samples with three optimal prediction coefficients, while the 1T1DF with a fractional pitch lag is constrained to use fixed interpolation coefficients.

Fig. 6 shows the performance of the 3T1DF switching configuration. With switching and sufficiently high interpolation ratio (more than 4), this configuration outperforms 3T3DF with $D = 1$. The cost of providing $D = 4$ for all pitch lags is 2 bits, while the cost of providing the two extra coefficients of a 3T3DF filter is certainly more than 2 bits. We can also compare two other cases, 3T1DF with switching ($D = 1$) and 1T1DF with $D = 2$. The cost of providing switching and interpolation are each 1 bit, but the 1T1DF with half sample lag resolution outperforms the switching case with no interpolation. However, as we allocate more bits to compare 3T1DF with switching and $D = 2$ with 1T1DF ($D = 4$), the performance is essentially the same. With another bit allocated (3T1DF with switching, $D = 4$ and 1T1DF with $D = 8$), the 3T1DF configuration pulls slightly ahead.

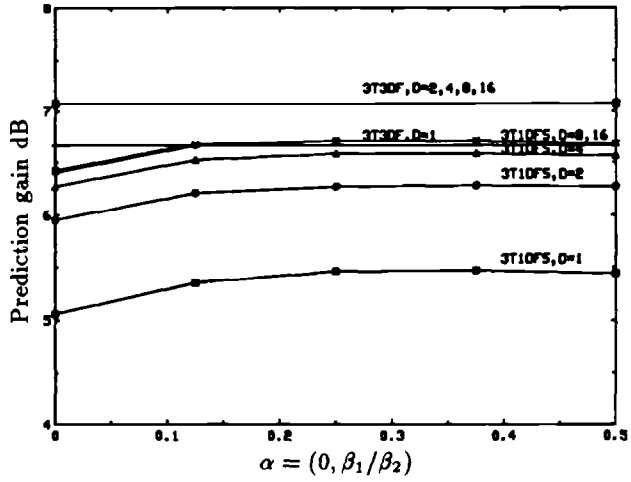

Fig. 6 The pitch prediction gain of a 3T1DF pitch filter with switching, male speech

The pitch prediction gain of 3T2DF filters is compared with 3T3DF and 1T1DF filters for different interpolation ratios in Fig. 7. The prediction gain for 3T2DF with a fractional pitch lag is close to that of the 3T3DF for both male and female speech. The 3T2DF performs better than the 3T1DF, since it always chooses an optimal α. But more interesting is that the 3T2DF filter with interpolation ratio at least 4, performs nearly as well as a 3T3DF filter with the same interpolation ratio.

5. Summary

We have presented two pseudo-three-tap pitch filter configurations, 3T2DF and 3T1DF, and derived the formulations of the optimal parameters for these pitch filters.

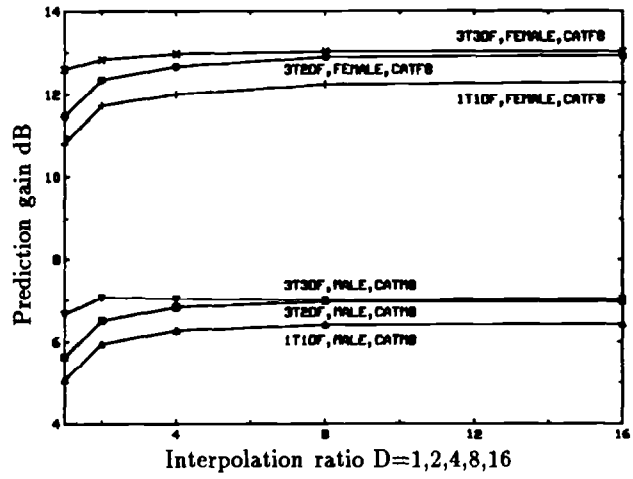

Fig. 7 The pitch prediction gain of a 3T2DF pitch filter for different value of D, male and female speech

The pseudo-three-tap pitch filter has fewer degrees of freedom than a traditional three-tap pitch filter, that is, fewer parameters need to be coded for transmission in a speech coding context. The 3T1DF is essentially a three-tap pitch filter with the first and third coefficients set to a fixed ratio of the second coefficient. A small but noticeable improvement of the 3T1DF is obtained compared to a one-tap pitch filter with no additional bit rate required. The 3T2DF can be considered as a 3T1DF with an adaptive optimal ratio of the outer and middle coefficients. The extra degree of freedom buys a better performance. A compromise is a switched 3T1DF configuration, with the switching costing only 1 extra bit.

When we compare configurations using interpolation, the results are more mixed. For the 3T1DF case, the improvement over 1T1DF is largest for $D = 1$ and decreasing for larger D. For the 3T1DF case with switching, the improvement over 1T1DF goes from a negative value to a positive value as D increases.

While the evaluations in this study have been in terms of prediction gains, we believe that the improved frequency response may be beneficial in low bit-rate speech coders to obtain a better reconstructed speech quality.

Reference

1. M.R. Schroeder and B.S. Atal,"Code-excited linear prediction (CELP): high quality speech at very low bit rates", *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Tampa, FL), March 26–29, 1985, pp. 937–940
2. V. Iyengar and P. Kabal,"A low delay 16 kbits/s speech coder", *IEEE Trans. Signal Processing*, Vol. 39, May 1991, pp. 1049–1057
3. J. Chen and M. Rauchwerk,"An 8 kb/s Low-Delay CELP Speech Coder," *IEEE Global Telecommun. Conf.* (Phoenix, AZ), Dec. 2–5, 1991, pp. 53.7.1–53.7.5
4. P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution", *IEEE Trans. on Signal Processing*, Vol. 39, Mar. 1991
5. R.E. Crochiere and L.R. Rabiner. *Multirate Digital Signal Processing.*, Prentice Hall, Englewood Cliffs, NJ, 1983