ELSEVIER

Speech Communication 14 (1994) 339–358

SPEECH COMMUNICATION

Pseudo-multi-tap pitch filters in a low bit-rate CELP speech coder *

Yasheng Qian [1], Gebrael Chahine and Peter Kabal *

Telecommunications and Signal Processing Laboratory, Department of Electrical Engineering, McGill University, 3480 University Street, Montréal, Québec, H3A 2A7, Canada

Received 28 February 1994; revised 9 June 1994

Abstract

The pitch filter in a low bit-rate CELP speech coder has a strong impact on the quality of the reconstructed speech. In this paper we propose a pseudo-multi-tap pitch filter with fewer degrees of freedom than the number of prediction coefficients, but which gives a higher pitch prediction gain and a more appropriate frequency response than a conventional one-tap pitch filter. First, we present an analysis model for the pseudo-multi-tap pitch prediction filter. Then, we introduce a pseudo-multi-tap pitch prediction filter with a fractional pitch lag. The prediction gain of the pseudo-multi-tap pitch filter is compared to that of conventional one-tap and three-tap pitch filters with integer and non-integer pitch lags. A switching configuration is also studied. This filter switches modes depending on the prediction gain. The stability of a pseudo-multi-tap pitch synthesis filter in a CELP coder is considered. We propose a stabilization method with a relaxed stability test. This relaxed test gives better results than a strict stability test. Finally, we have incorporated the pseudo-multi-tap pitch filter into a 4.8 kbit/s CELP speech coder. Both the objective SNR and subjective quality are better than for a conventional one-tap pitch filter.

Zusammenfassung

Das Sprachgrundfrequenzfilter in einem CELP-Sprachcoder mit geringer Bitrate übt einen starken Einfluß auf die rekonstruierte Sprache aus. In diesem Artikel schlagen wir ein *pseudo-multi-tap* (pseudo Polykoeffizienten) Sprachgrundfrequenzfilter mit einem geringeren Freiheitsgrad als der Anzahl der Prädiktionskoeffizienten entspricht vor, das aber einen höheren Langzeitprädiktionsgewinn und eine passendere Frequenzantwort aufweist, als ein herkömmliches Prädiktionsfilter mit einem einzigen Koeffizienten. Wir stellen ein Analysemodel für das pseudo-multi-tap Sprachgrundfrequenzfilter vor, mit einer im Vergleich zur Grundfrequenz sehr kleinen Schrittweitencodierung. Der Prädiktionsgewinn des pseudo-multi-tap Sprachgrundfrequenzfilters wird mit dem von herkömmlichen Einkoeffizienten- und Dreikoeffizientenfiltern verglichen; die Schrittweiten sind in Bezug auf die Sprachgrundfrequenz sowohl ganzzahlig als auch nicht ganzzahlig codiert. Es wird die wechselweise Verwendung beider Modi in Abhängigkeit vom Prädiktionsgewinn untersucht. Die Stabilität des pseudo-multi-tap Synthesefilters

* Corresponding author. Tel: (514) 398 7130; Fax: (514) 398 4470; Email: kabal@tsp.ee.mcgill.ca

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

[1] Yasheng Qian is on leave from Tsinghua University, Beijing, China

SSDI 0167-6393(94)00014-2

in einem CELP-Coder wird mit in Betracht gezogen. Wir schlagen eine Stabilisierungsmethode mit vermindertem Stabilitätstest vor. Diese liefert bessere Resultate als der strenge Stabilitätstest. Schließlich haben wir diese pseudo-multi-tap Sprachgrundfrequenzfilter in einem 4.8 kbit/s CELP-Sprachcoder eingebaut. Sowohl objektive SNR als auch subjektive Qualitätsbeurteilung sind besser als in einem herkömmlichen Einkoeffizienten-Langzeitprädiktionsfilter.

Resumé

Le filtre à long-terme d'un codeur de parole CELP à bas-débit a une influence notable sur la qualité de la parole reconstruite. Dans cet article, nous proposons un pseudo-filtre de prédiction à long terme à plusieurs coefficients qui possède moins de degrés de liberté que le nombre de coefficients de prédiction, mais donne un meilleur gain en prédiction à long-terme et une meilleure réponse en fréquence qu'un filtre de prédiction conventionnel à un seul coefficient. Nous proposons d'abord un pseudo-filtre de prédiction à long-terme à plusieurs coefficients à décodage fractionnel par rapport à la fréquence fondamentale. Le gain de prédiction de ce filtre est comparé à celui des filtres classiques à un seul coefficient et à trois coefficients, avec des décalages entiers et fractionnaires. Nous décrivons aussi une configuration autorisant leur commutation, celle-ci étant commandée par le gain de prédiction. La stabilité du filtre lors de la phase de synthèse est étudiée pour un codeur CELP: nous décrivons une méthode de stabilisation comprenant un test de stabilité affaiblie. Elle donne de meilleurs résultats qu'un test de stabilité stricte. Pour finir, nous avons incorporé notre pseudo-filtre de prédiction à long terme à plusieurs coefficients dans un codeur CELP à 4.8 kbit/s. Tant le rapport objectif signal à bruit que la qualité subjective ont été améliorés par rapport à ceux mesurés avec un filtre de prédiction à long terme conventionnel à un seul coefficient.

Key words: Speech coding; Pitch filter; Prediction gain; Fractional pitch lag; Stability

1. Introduction

A pitch filter cascaded with a formant filter is widely employed in many low bit-rate code-excited linear prediction (CELP) speech coders (Schroeder and Atal, 1989; Iyengar and Kabal, 1991; Campbell et al., 1990; Ramachandran and Kabal, 1989). In many cases, a one-tap pitch filter with integer or non-integer pitch lags is used. A three-tap pitch prediction filter provides a higher prediction gain than a one-tap pitch prediction filter. However, additional bits are required to adequately code the pitch filter coefficients.

The objective of our study is to develop a more efficient way to represent a multi-tap pitch filter in a low rate speech coder. There are two kinds of pitch filters. The pitch filter at the analysis stage of a speech coder is a non-recursive pitch prediction filter. The pitch filter used at the synthesis stage of a speech coder is the inverse filter to the pitch prediction filter, i.e., a recursive filter. The placements of the pitch prediction filter and the pitch synthesis filter are shown in Fig. 1. In practice, if an analysis-by-synthesis procedure is used, the synthesis filter is included in a closed loop search.

The frequency response of a one-tap pitch synthesis filter shows a constant envelope (see Fig. 2). The search for pseudo-multi-tap pitch filters was motivated by the observation that the spectrum of a conventional three-tap pitch filter often shows a diminishing envelope with increasing frequency in some voiced segments (see Fig. 3). This corresponds to a large center coefficient and smaller side coefficients. Such a frequency response adds more pitch structure at low frequencies than at high frequencies. Consider the case of an integer lag, one-tap pitch filter. Suppose that the true pitch lag is in-between integer values. The frequency response of an integer lag filter will be up to 90 degree out of phase at the half-sampling frequency. At low frequencies such fractional lag errors do not affect the spectral fit. One effect of a shaped envelope such as provided by a multi-tap pitch filter, is that the effect of mismatches at high frequencies can be deemphasized. It should, however, be noted that the multi-tap pitch filter can also exhibit other spectral envelope (see Figs. 11–13 later).

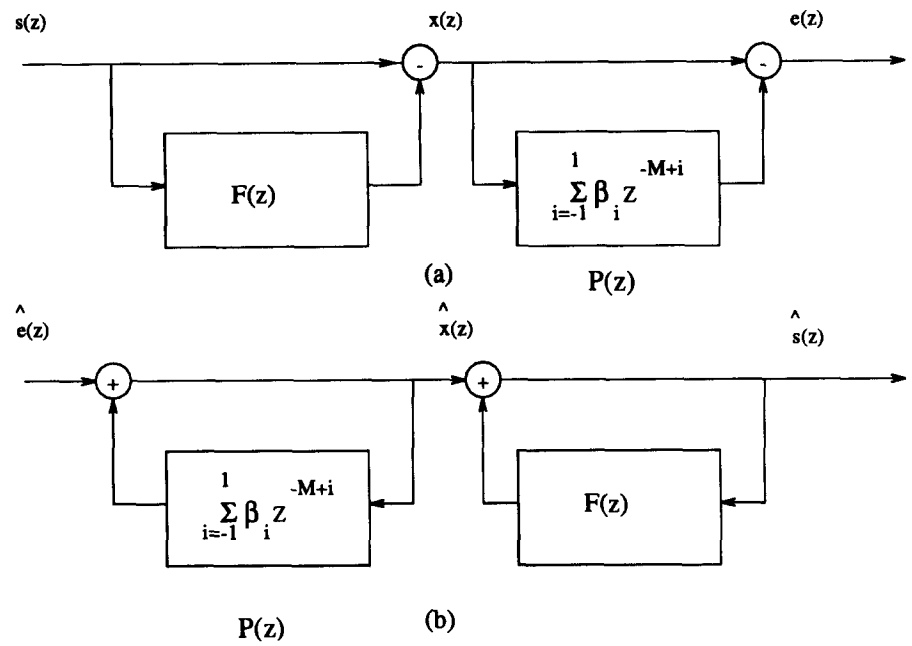

Fig. 1. Block diagram of a pitch filter cascaded with a formant filter $F(z)$: (a) pitch prediction filter, (b) pitch synthesis filter.

One 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 explicit (Kroon and Atal, 1991). Additional bits are needed to code the resulting higher resolution pitch lags. However, such one-tap fractional-pitch filters still have a constant envelope frequency response. Fractional-lag filters solve the interpolation problem without addressing the spectral envelope issue.

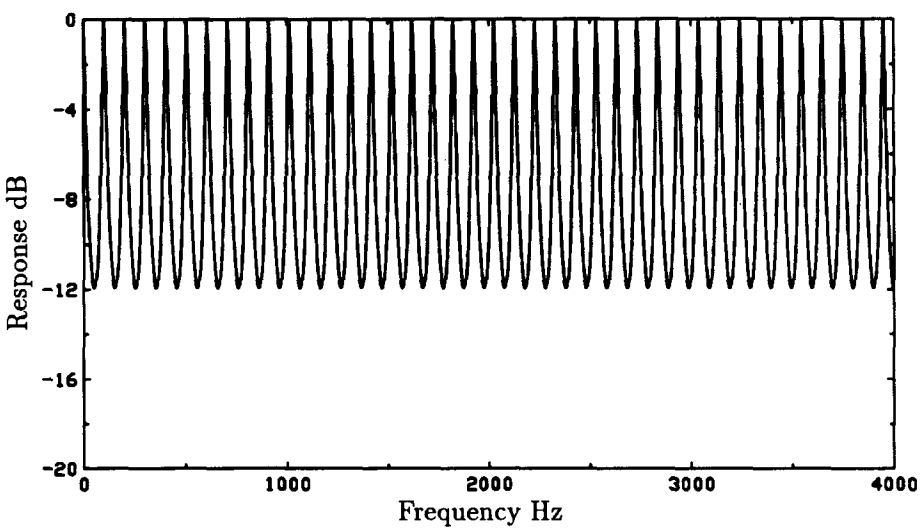

Fig. 2. Frequency response of a one-tap pitch synthesis filter.

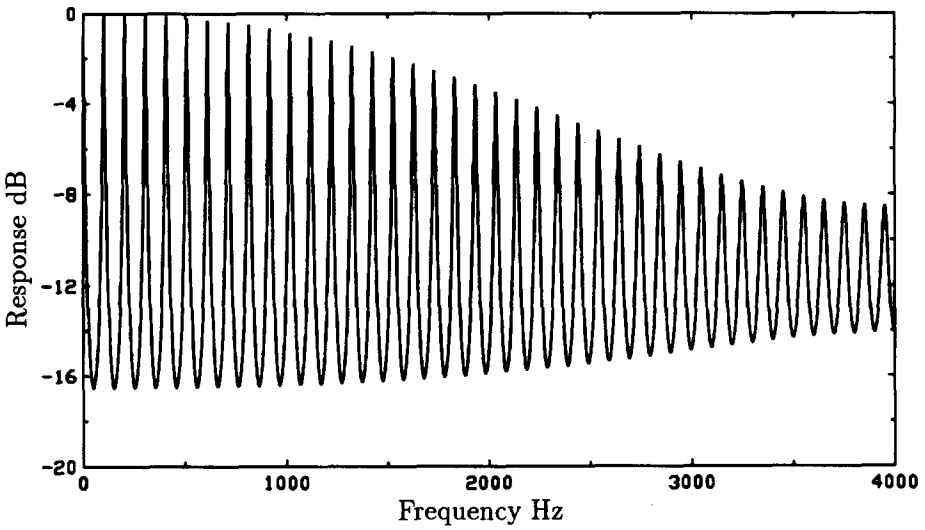

Fig. 3. Frequency response of a three-tap pitch synthesis filter with coefficients (0.27, 0.52, −0.06).

The stability of a pitch synthesis filter is another important issue in a CELP coder. Since the pitch synthesis filter is recursive and is usually determined by a covariance method, it can result in an unstable pitch filter. In practice, unstable pitch filters can greatly degrade the reconstructed speech quality. This problem along with several stabilization methods has been studied in (Ramachandran and Kabal, 1987) by analyzing the original speech. However, the pitch filter parameters are determined by an analysis-by-synthesis search procedure in a CELP coder. Although the effect of the noise enhancement in an unstable pitch filter is taken into account in a closed-loop search algorithm, unstable pitch filters can still impair speech quality.

We analyze the effect of stability of the pseudo-multi-tap pitch filter. Then, we present stabilization methods for pseudo-multi-tap pitch filters to improve the speech coder quality.

In this paper, we first focus on a general analysis model for the pseudo-multi-tap pitch prediction filter. Then, we describe the pseudo-multi-tap pitch filter with a fractional pitch lag. The pitch prediction gains of the pseudo-multi-tap pitch filters are compared to conventional one-tap and three-tap pitch predictors with integer or non-integer pitch lags. A switching configuration is also explored. The frequency response of the pseudo-multi-tap pitch filter is examined and the stability of such filters is considered. A stabilization procedure with a relaxed stability check is proposed. Finally, we present the performance of a 4.8 kbit/s CELP coder with different filter configurations of pseudo-multi-tap pitch filters.

2. A pseudo-multi-tap pitch prediction filter

A pseudo-multi-tap pitch filter is an n-tap pitch prediction filter which has fewer than n degrees of freedom. We illustrate a pseudo-multi-tap filter with a three-tap example. Let a traditional three-tap pitch prediction filter have three non-zero coefficients at lags $M-1$, M, $M+1$, with M representing the pitch lag. Let the three non-zero coefficients of the three-tap pitch filter be β_{-1}, β_0 and β_{+1}. This gives three degrees of freedom. We can restrict this filter to two degrees of freedom, while maintaining a symmetrical set of coefficients, by assigning

$$\beta_{-1}=\beta_{+1}=\gamma, \quad \beta_0=\beta. \tag{1}$$

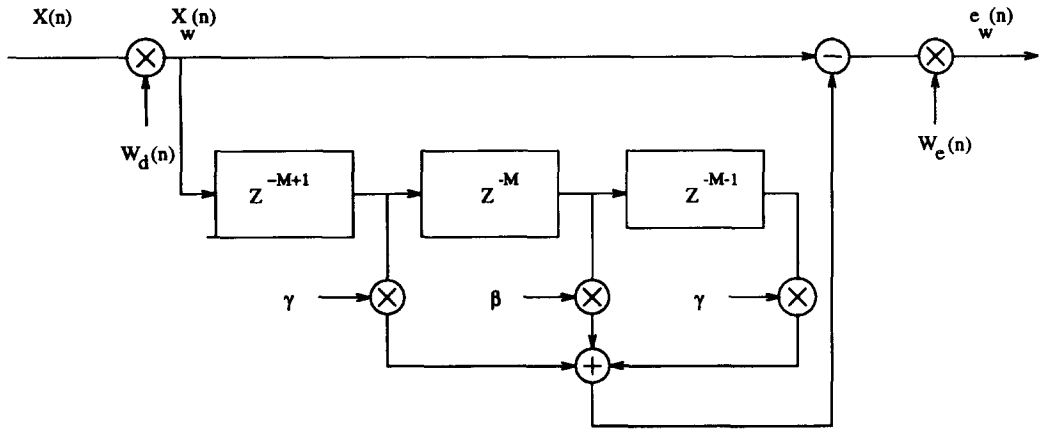

Fig. 4. Analysis model for a pseudo-multi-tap predictor.

Both β and γ can be chosen to give the best performance. We can further restrict the pseudo-multi-tap filter to one degree of freedom by letting $\gamma = \alpha\beta$ with a fixed value of α. The notation for pseudo-multi-tap filters are nTmDF, meaning n taps, m degrees of freedom. Thus, a conventional three-tap filter is 3T3DF (β_{-1}, β_0 and β_{+1} variable). The pseudo-three-tap filters are 3T2DF (γ and β variable) and 3T1DF (α fixed, β variable).

An analysis model for calculating the prediction coefficients of the pseudo-multi-tap pitch predictor with a transversal implementation is shown in Fig. 4. 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

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

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

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

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 $0 \leqslant n \leqslant L-1$. The lag M is chosen as that which is optimal for a one-tap pitch predictor (Ramachandran and Kabal, 1989). For the case of 3T2DF, the coefficients β and γ are computed by minimizing ε^2. The minimization of ε^2, setting partial derivatives of ε^2 to zero, leads to a set of linear equations which can be written in matrix form,

$$\begin{bmatrix} A & B \\ B & D \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} = \begin{bmatrix} E \\ F \end{bmatrix}, \tag{4}$$

where

$$\begin{aligned} A &= \phi(M-1, M-1) + \phi(M+1, M+1) + 2\phi(M-1, M+1), \\ B &= \phi(M-1, M) + \phi(M, M+1), \qquad D = \phi(M, M), \\ E &= \phi(0, M-1) + \phi(0, M+1), \qquad F = \phi(0, M), \end{aligned} \tag{5}$$

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

$$\phi(i,j) = \sum_{n=-\infty}^{\infty} w_e^2(n) x_w(n-i) x_w(n-j), \tag{6}$$

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

$$\beta_{\text{opt}} = (AF - BE)/(AD - B^2), \qquad \gamma_{\text{opt}} = (DE - BF)/(AD - B^2). \tag{7}$$

For the 3T1DF case, β is

$$\beta_{\text{opt}} = \frac{\alpha\phi(0, M-1) + \phi(0, M) + \alpha\phi(0, M+1)}{\alpha^2\phi_3 + \phi(M, M) + 2\alpha\phi_2}, \tag{8}$$

where

$$\begin{aligned} \phi_3 &= \phi(M-1, M-1) + 2\phi(M-1, M+1) + \phi(M+1, M+1), \\ \phi_2 &= \phi(M-1, M) + \phi(M, M+1). \end{aligned} \tag{9}$$

3. A fractional pseudo-multi-tap pitch prediction filter

The use of a fractional pitch lag has proved to be an accurate and efficient means to characterize speech periodicity in low bit-rate speech coders. Fractional pitch lags can also be used in pseudo-multi-tap pitch prediction filters. A non-integer pitch lag can be expressed as an integer number of samples plus a fraction. Let the pitch resolution be $1/D$. The fractional part of the pitch lag can be expressed as l/D, where $l = 0, 1, \ldots, D-1$ $(0 \leqslant l/D \leqslant 0)$. The pseudo-multi-tap filter then acts on the interpolated samples, denoted by $x_w^{(l)}(n-(M-1))$, $x_w^{(l)}(n-M)$, $x_w^{(l)}(n-(M+1))$. The fractional delay is implemented using an interpolation filter. This filter delays the signal at the higher rate by an integer number of samples. The subsampled output of this filter is the desired fractionally delayed signal. Note that usually the filter coefficients are chosen to give $x_w^{(0)}(n) = x_w(n)$.

A polyphase filter structure (Crochiere and Rabiner, 1983) can be used to obtain the interpolated samples. For each fractional phase l, the impulse response $p^{(l)}(n)$ of the polyphase filter is obtained by sub-sampling an appropriate interpolating filter. In our case, we use an interpolation filter which is a Hamming-windowed ideal low-pass filter,

$$p^{(l)}(n) = w_h(n - l/D) \frac{\sin(\pi(n - l/D))}{\pi(n - l/D)}, \tag{10}$$

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

We have chosen to have the same number of coefficients $2I$ for each of the polyphase component filters $(l \neq 0)$. The resulting value which corresponds to the interpolated sample at $n + l/D$ is given by

$$x_w^{(l)}(n) = \sum_{k=0}^{2I-1} p^{(l)}(k - I) x_w(n-k), \tag{11}$$

where $2I$ is the number of the coefficients of the polyphase filter. The delay of the causal interpolation filter at the original sampling rate is I. The prediction error for a (fractional) pitch lag of $M - l/D$ can be written as

$$e(n) = x_w(n) - \sum_{i=-1}^{1} \sum_{k=0}^{2I-1} \beta_i p^{(l)}(k - I) x_w(n - (M+i) - k). \tag{12}$$

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 functions with fractional delays are
For $i \neq 0$,

$$\phi^{(l)}(i,j) = \sum_{n=-\infty}^{\infty} \sum_{k=0}^{2I-1} p^{(l)}(k-I)x_w(n-i-k) \sum_{k=0}^{2I-1} p^{(l)}(k-I)x_w(n-j-k); \tag{13}$$

For $i = 0$,

$$\phi^{(l)}(0,j) = \sum_{n=-\infty}^{\infty} x_w(n) \sum_{k=0}^{2I-1} p^{(l)}(k-I)x_w(n-j-k). \tag{14}$$

For each $\phi^{(l)}(i, j)$ and $\phi^{(l)}(0, j)$, $l \neq 0$, we have to convolve the impulse response of the polyphase filter and the weighted input samples to get the corresponding interpolated samples. In fact, each interpolated sample has to be manipulated many times to determine the best pitch lag $M - l/D$ and the optimal prediction coefficients. The computation load is reduced if we first calculate the interpolated samples $x_w^{(l)}(n)$, for $l = 1, \ldots, D-1$. Then, (13) and (14) become

$$\phi^{(l)}(i,j) = \sum_{n=-\infty}^{\infty} x_w^{(l)}(n-i)x_w^{(l)}(n-j), \qquad \phi^{(l)}(0,j) = \sum_{n=-\infty}^{\infty} x_w^{(l)}(n)x_w^{(l)}(n-j). \tag{15}$$

In order to identify the pitch lag $M - l/D$, we find the lag which minimizes the error for a one-tap pitch filter. We obtain the minimum square error $(\varepsilon^2)^{(M-l/D)}$ corresponding to the optimal prediction coefficient β_{opt} for the given delay $M - l/D$.

$$(\varepsilon^2)^{(M-l/D)} = \sum_{n=-\infty}^{\infty} w_e(n)^2 \left[x_w(n)^2 - \frac{(\phi^{(l)}(0,M))^2}{(\phi^{(l)}(M,M))^2} \right]. \tag{16}$$

The minimum of $(\varepsilon^2)^{(M-l/D)}$ corresponds to the maximum of $(\phi^{(l)}(0, M))^2/(\phi^{(l)}(M, M))^2$ over the range of allowable pitch lags.

4. Pitch prediction gain

The pitch prediction gain is used to compare the performance of the pseudo-multi-tap pitch filters to conventional one-tap and three-tap pitch filters. The predictor gain PG (expressed in dB) is the ratio of the energy at the input of the predictor to that of the prediction error,

$$\text{PG}_{\text{dB}} = 10 \log \frac{\sum_{n=-\infty}^{\infty} x_w^2(n)}{\varepsilon^2}. \tag{17}$$

In all cases, the pitch prediction filters are 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 is that which is best for a one-tap pitch filter.

Table 1 shows the average pitch prediction gains for a number of configurations, all with integer pitch lags. The results are shown for a single sentence. Note that the performance of the 3T1DF configuration depends on the value of α chosen. The results shown in the table indicates that $\alpha = 0.125$ is good for both male and female speech. The average gains are about 0.2 dB higher than a conventional one-tap

Table 1
Pitch prediction gains in dB. The notation, nTmDF, means n taps, m degrees of freedom, integer lags. Note that a 3T1DF with $\alpha = 0$ is in fact a 1T1DF filter

Type	α	Prediction gain (dB)	
		male	female
1T1DF	0.000	5.06	10.81
3T1DF	0.125	5.24	11.06
3T1DF	0.25	5.21	10.78
3T1DF	0.375	5.06	10.35
3T1DF	0.5	4.87	9.95
3T2DF	—	5.62	11.48
3T3DF	—	6.66	12.60

pitch filter. The 3T2DF filter is about 0.6 dB better than a one-tap filter and the 3T3DF filter is about 1.6 dB better than a one-tap filter.

Next we compare the pseudo-three-tap pitch filter with a fractional pitch lag to one-tap and 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 (maximum 16) were used. 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.

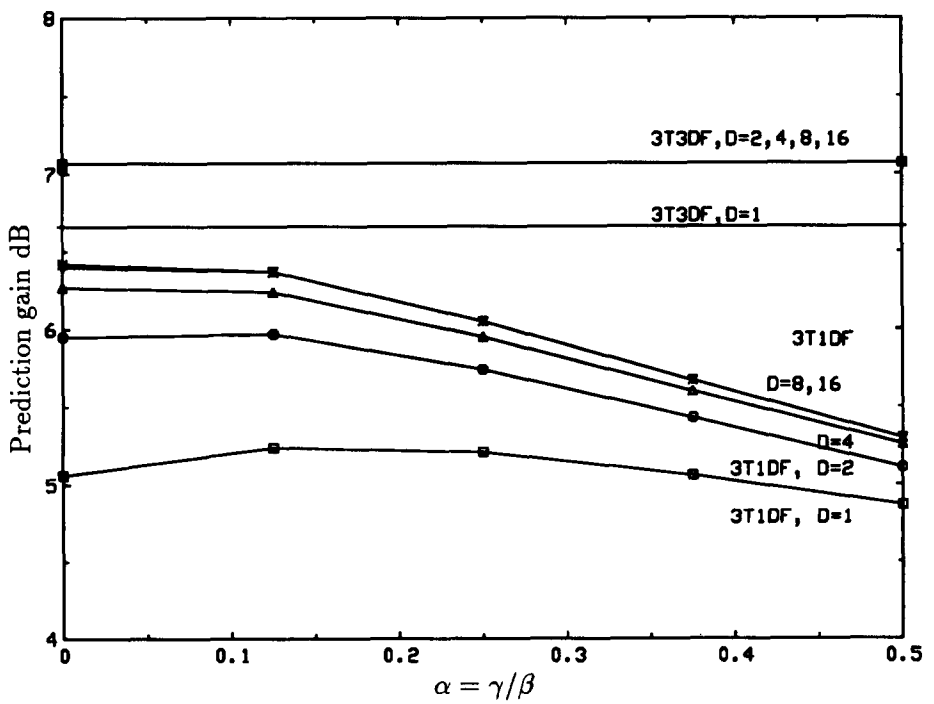

Fig. 5. Pitch prediction gains for pitch filters versus α for different values of D, male speech.

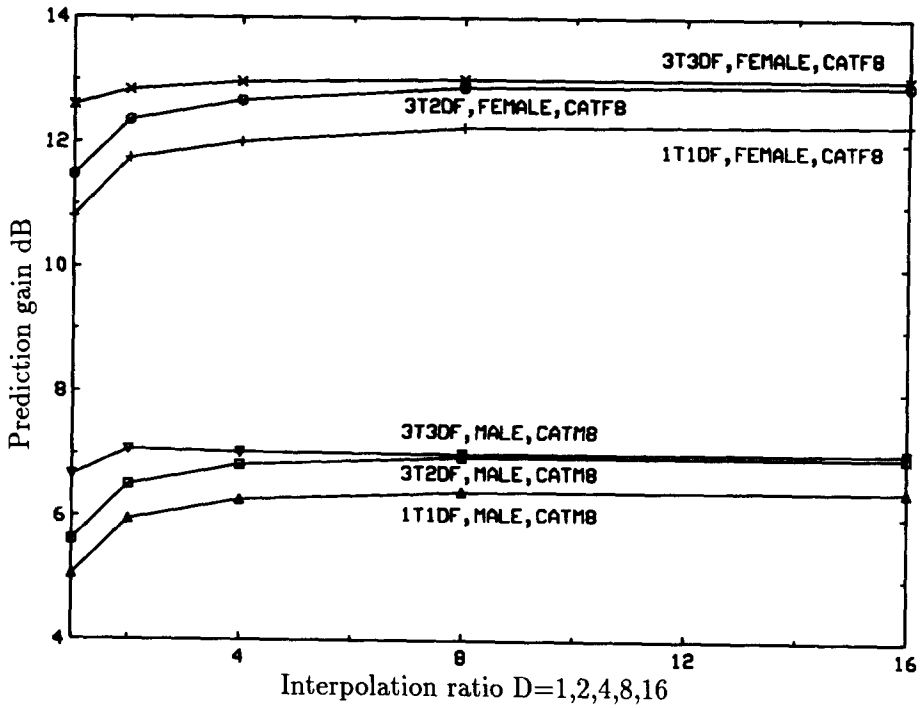

Fig. 6. The pitch prediction gain of a 3T2DF pitch filter for different value of D, male and female speech.

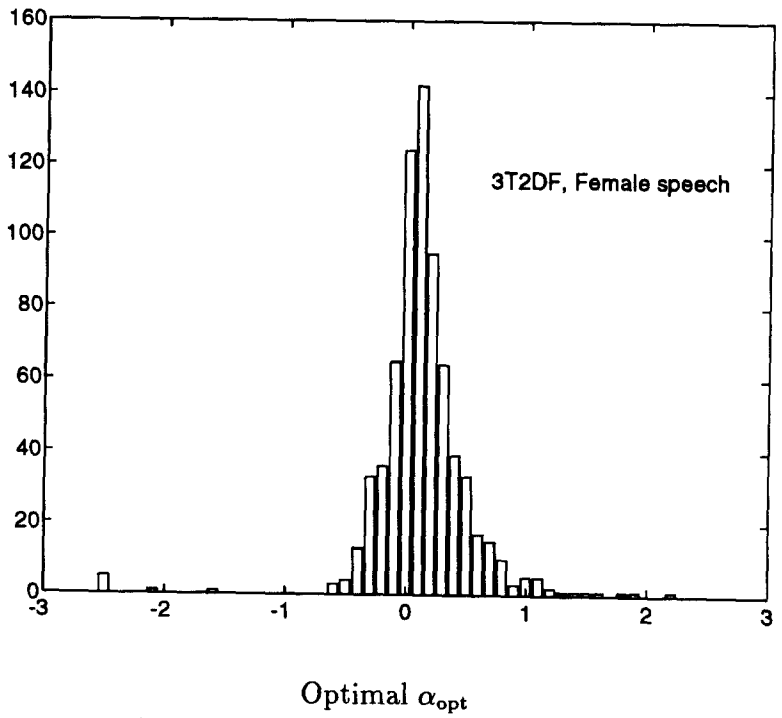

Fig. 7. The histogram of the optimal ratio α_{opt} of 3T2DF.

We have also evaluated a conventional three-tap pitch filter 3T3DF with a fractional pitch lag. The 3T3DF filter with an interpolation ratio of $D = 2$ gives an increased prediction gain of 0.41 dB for male speech. The 3T3DF filter with higher interpolation ratios $D > 4$ 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 the 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.

The pitch prediction gain of 3T2DF filters is compared with 3T3DF and 1T1DF filters for different interpolation ratios in Fig. 6. 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 filter performs better than the 3T1DF filter, 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.

The 3T2DF filter can be viewed as a special case of 3T1DF filter with an optimum α_{opt}. We obtain the optimum value of the α_{opt} from (7),

$$\alpha_{\text{opt}} = \frac{\gamma_{\text{opt}}}{\beta_{\text{opt}}} = \frac{DE - BF}{AF - BE}. \tag{18}$$

The histogram of the α_{opt} of the 3T2DF for a female speech is given in Fig. 7. It shows that the mean value of α_{opt} is 0.127 and the median value of α_{opt} is 0.108. The corresponding values for the male speech during voiced segments are 0.131 and 0.110, respectively. These results justify the use of α equal to 0.125 as a reasonable choice for the 3T1DF filter.

5. Switching configuration

We have found that the pitch prediction gain of the 3T1DF pitch filter is higher than that of the 1T1DF configuration by 1.5–2.0 dB in some speech frames, but in others it can in fact be slightly worse than 1T1DF. This suggests that it is possible to combine these two configurations, switching to the one which performs the best.

The minimum mean square prediction error $\varepsilon^2_{\text{min}}$ can be obtained by substituting the optimum pitch prediction coefficient β_{opt} (8) into (3).

For the 3T1DF case,

$$\varepsilon^2_{\text{min}\,|\,3\text{T1DF}} = \phi(0,0)[1 - E_N(M,\alpha)],$$

where

$$E_N(M,\alpha) = \frac{[\alpha\phi(0,M-1) + \phi(0,M) + \alpha\phi(0,M+1)]^2}{[2\alpha\phi(M-1,M) + \phi(M,M) + 2\alpha\phi(M,M+1) + \phi_S]\phi(0,0)}, \tag{19}$$

$$\phi_S = \alpha^2[\phi(M+1,M+1) + \phi(M-1,M-1) + 2\phi(M-1,M+1)]$$

For the 1T1DF case, as a special case of 3T1DF with $\alpha = 0$,

$$\varepsilon^2_{\text{min}\,|\,1\text{T1DF}} = \phi(0,0)[1 - E_N(M,0)], \quad E_N(M,0) = \frac{\phi(0,M)^2}{\phi(M,M)\phi(0,0)}. \tag{20}$$

Thus, if $\text{PG}_{3\text{T1DF}} > \text{PG}_{1\text{T1DF}}$,

$$E_N(M,\alpha) > E_N(M,0). \tag{21}$$

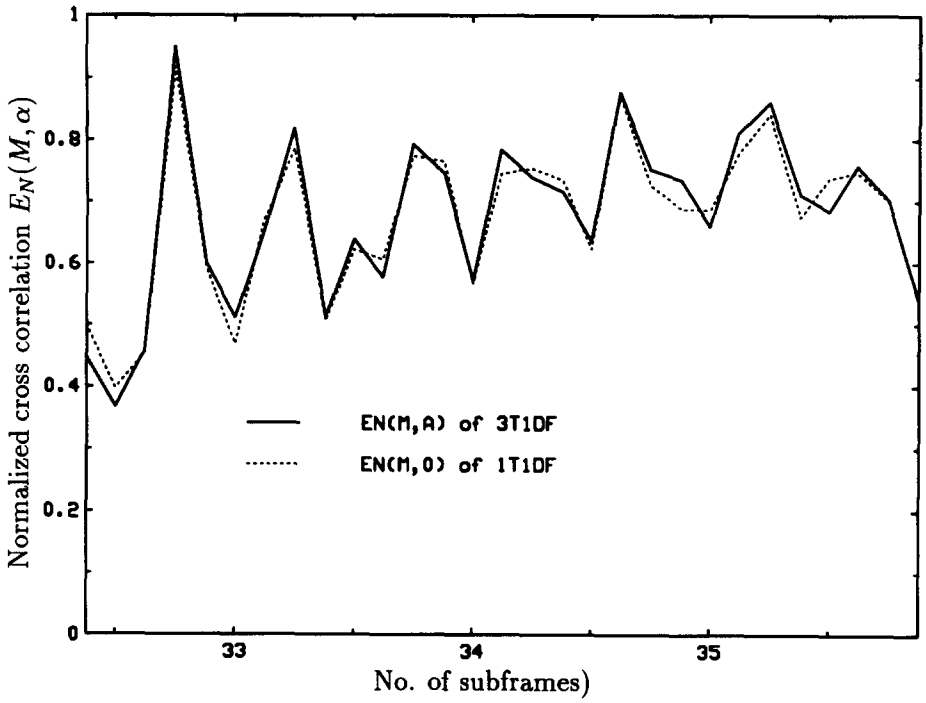

Fig. 8. The normalized cross-correlation $E_N(M, \alpha)$ of a 3T1DF and 1T1DF, $\alpha = 0.125$.

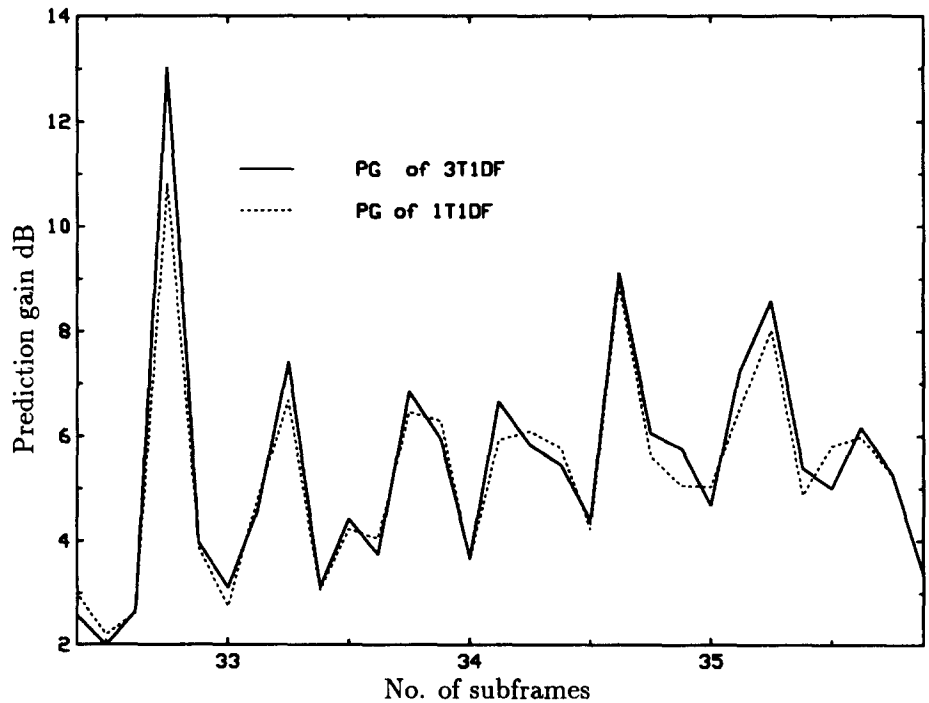

Fig. 9. The pitch prediction gain of a 3T1DF and 1T1DF pitch filter.

Table 2
Pitch prediction gains in dB. The two values in the second column indicate that α switches between these values

Type	α	Prediction gain (dB)	
		male	female
1T1DF	0, 0.0	5.06	10.81
3T1DF	0, 0.125	5.35	11.22
3T1DF	0, 0.25	5.46	11.25
3T1DF	0, 0.375	5.47	11.20
3T1DF	0, 0.5	5.44	11.18

Most segments of a speech signal meet the condition (21). Therefore, the average PG_{3T1DF} is higher than the PG_{1T1DF}, as shown in Table 1. However, there are some subframes which do not conform to the condition. Fig. 8 shows the normalized cross correlation $E_N(M, \alpha)$ for a 3T1DF filter and $E_N(M, 0)$ for a 1T1DF filter in different subframes. Fig. 9 shows the pitch prediction gain for 3T1DF and 1T1DF filters. $E_N(M, \alpha)$ is lower than $E_N(M, 0)$ in several subframes. Therefore, the pitch prediction gain PG_{3T1DF} of these corresponding subframes is lower than the PG_{1T1DF}.

In the switching configuration we select the pitch prediction filter, 3T1DF or 1T1DF, whichever has the higher pitch prediction. Table 2 and Fig. 10 show the results (3T1DFS, $D = 1$) of switching between $\alpha = 0$ (the 1T1DF case), and another non-zero value. With switching, $\alpha = 0.250$ is preferable to $\alpha = 0.125$. Note that switching between α's uses one bit of information. This approach can also be considered to coarsely quantize the α parameter of a 3T2DF configuration.

Fig. 10 shows the performance of the 3T1DF switching configuration. With switching and sufficiently

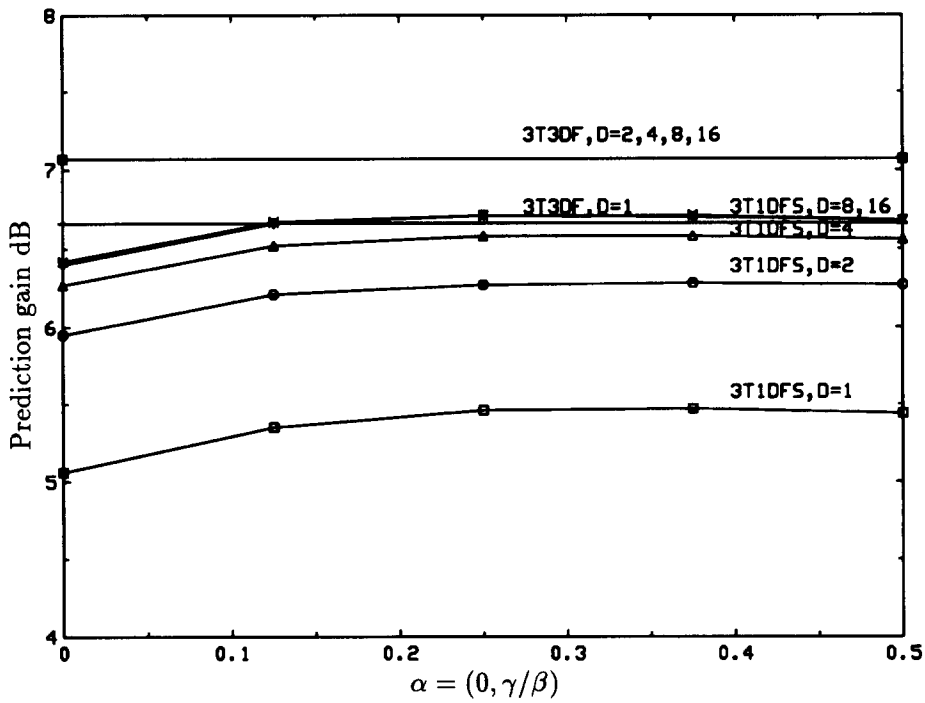

Fig. 10. The pitch prediction gain of a 3T1DF pitch filter with switching, male speech.

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.

6. Frequency response

The frequency response of the pitch synthesis filter affects the reconstructed speech spectrum in a CELP coder. We compare the frequency response of pseudo-multi-tap synthesis filters 3T2DF and 3T1DF with conventional 1T1DF and 3T3DF filters.

The frequency response of a 3T3DF pitch synthesis filter can be expressed as

$$H(e^{j\omega}) = \frac{1}{1 - \beta_{-1} e^{-j\omega(M-1)} - \beta_0 e^{-j\omega M} - \beta_{+1} e^{-j\omega(M+1)}}. \tag{22}$$

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

$$|H(e^{j\omega})| = \frac{1}{\sqrt{[\cos(\omega M) - \beta_0 - (\beta_{-1} + \beta_{+1})\cos(\omega)]^2 + [(\beta_{+1} - \beta_{-1})\sin(\omega) + \sin(\omega M)]^2}}. \tag{23}$$

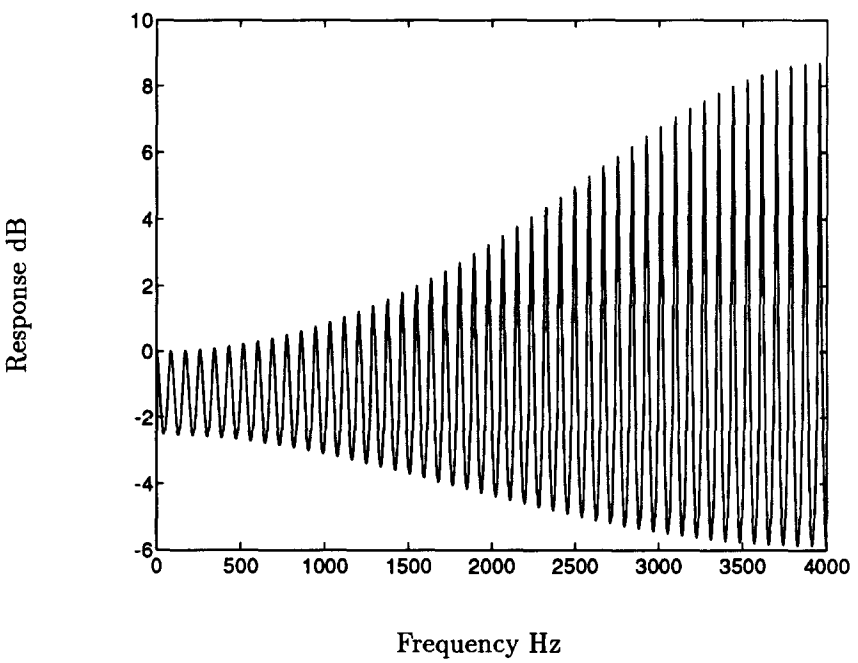

Fig. 11. Frequency responses of a three-tap pitch filter with coefficients (-0.14, 0.41, -0.14).

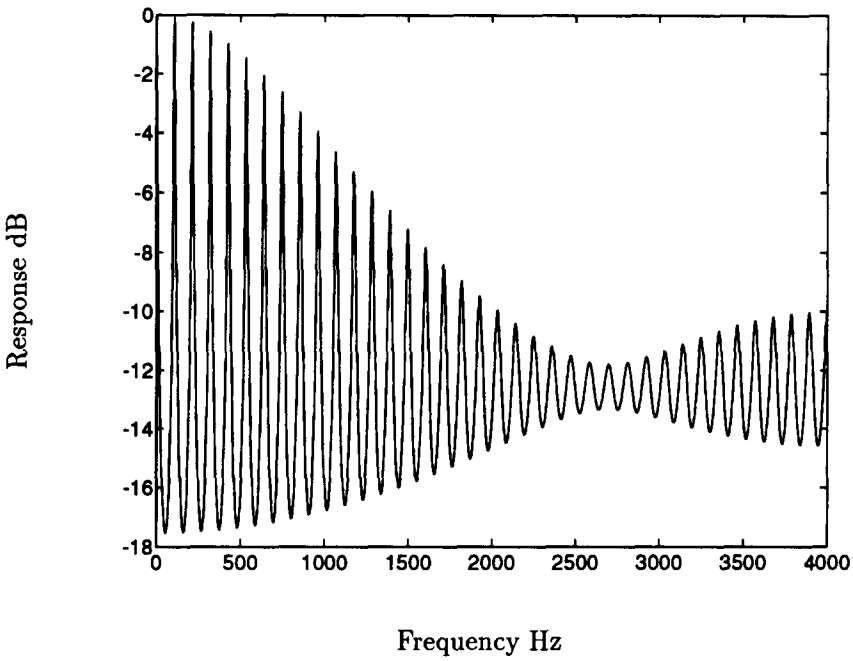

Fig. 12. Frequency responses of a three-tap pitch filter with coefficients (0.31, 0.25, 0.20).

Since the pitch period M is in the range of 20–147 (8 kHz sampling), we consider $M \gg 1$. The terms of $\cos(\omega M)$ and $\sin(\omega M)$ produce the quasi-harmonic structure in the frequency response. The envelope of the frequency response depends mainly 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 for $\omega = (0, \pi)$. The term $(\beta_{+1}-\beta_{-1})\sin(\omega)$ reaches a maximum of $(\beta_{+1}-\beta_{-1})$ at $\omega = \pi/2$. For a given pitch period M, the envelope depends on the values of β_{-1}, β_0, β_{+1}. There are four possible envelopes:

1. A decreasing monotonic shape, if $\beta_0 > (\beta_{-1}+\beta_{+1}) > 0$, as shown in Fig. 3. The term $(\beta_{-1}+\beta_{+1})\cos(\omega)$ decreases monotonically. Thus, the envelope of $|H(e^{j\omega})|$ also decreases with ω in (23).
2. An increasing monotonic envelope, if $(\beta_{-1}+\beta_{+1}) < 0$ and $|\beta_{-1}+\beta_{+1}| \approx \beta_0$, as shown in Fig. 11. Since the term $(\beta_{-1}+\beta_{+1})\cos(\omega)$ increases monotonically, the envelope of $|H(e^{j\omega})|$ increases in (23).
3. Two resonances, if $(\beta_{-1}+\beta_{+1}) > \beta_0 > 0$ and $|\beta_{+1}-\beta_{-1}| \gg 0$, as shown in Fig. 12. The term $(\beta_{+1}-\beta_{-1})\sin(\omega)$ makes an important contribution in the middle region. Since this term vanishes at $\omega = 0$ and at $\omega = \pi$, there is a valley in the middle region.
4. A resonance in the middle, if β_{-1} and β_{+1} have different signs, as shown in Fig. 13. The term $(\beta_{+1}-\beta_{-1})\sin(\omega)$ makes more contribution than the term $(\beta_{+1}+\beta_{-1})\cos(\omega)$.

For the case of a 3T2DF filter, $|H(e^{j\omega})|$ reduces to

$$|H(e^{j\omega})| = \frac{1}{\sqrt{[\cos(\omega M) - \beta - 2\gamma\cos(\omega)]^2 + [\sin(\omega M)]^2}}. \tag{24}$$

The amplitude of $|H(e^{j\omega})|$ has two possible envelopes: a decreasing envelope if γ has the same sign as β; an increasing envelope if γ has a large value with a different sign.

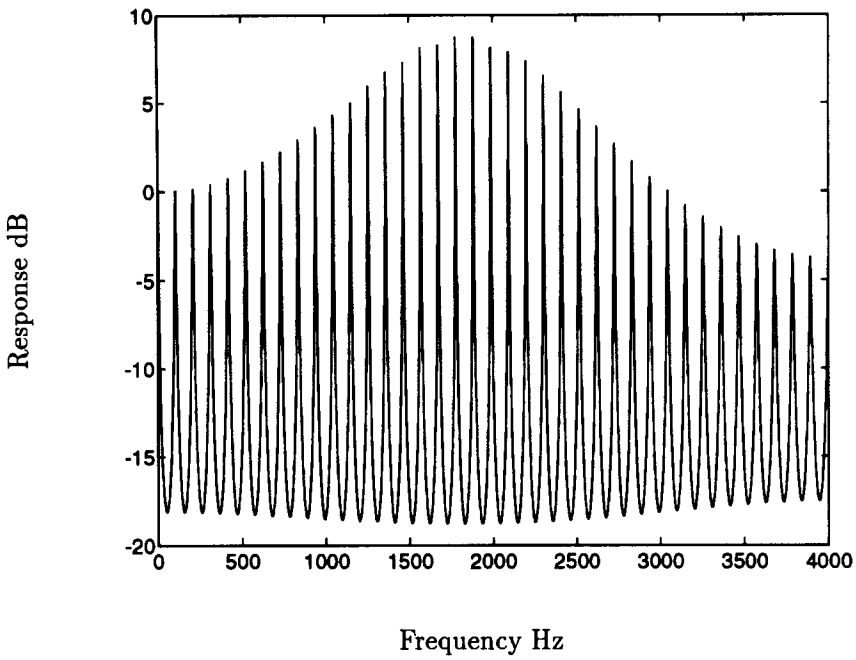

Fig. 13. Frequency responses of a three-tap pitch filter with coefficients (-0.25, 0.72, 0.32).

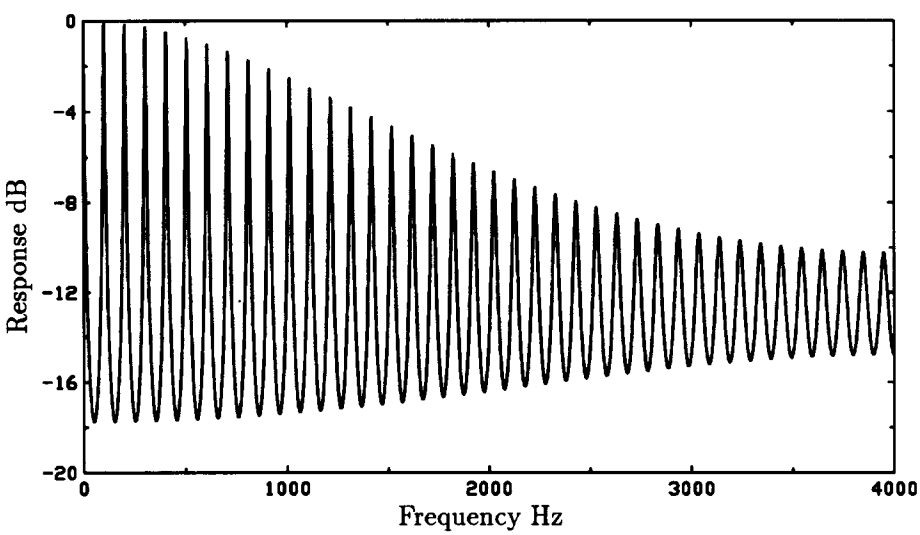

Fig. 14. Frequency response of a 3T1DF pitch synthesis filter with $\alpha = 0.25$, male speech.

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

$$|H(e^{j\omega})| = \frac{1}{\sqrt{[\cos(\omega M) - \beta(1 + 2\alpha\cos(\omega))]^2 + [\sin(\omega M)]^2}}. \tag{25}$$

The amplitude of $|H(e^{j\omega})|$ of a 3T1DF filter has a decreasing envelope for positive α. The frequency response of the 3T1DF filter with $\alpha = 0.25$ is shown in Fig. 14. This can be compared to Figs. 2 and 3. Let $\alpha = 0$ in (25). Then, $|H(e^{j\omega})|$ becomes a constant envelope of a 1T1DF pitch filter,

$$|H(e^{j\omega})| = \frac{1}{\sqrt{1 - 2\beta\cos(\omega M) + \beta^2}}. \tag{26}$$

7. Stability

In this section, we discuss the effect of an unstable pitch filter in a CELP coder. There are three procedures to determine the pitch filter parameters, the pitch lag M and the prediction coefficients $\{\beta_i\}$, $i = 1, 2, 3$: (1) Analyzing the original speech signal by solving a covariance matrix equation, as for a pitch prediction filter in Section 2 and in (Ramachandran and Kabal, 1987); (2) Jointly optimizing the excitation codebook index i, the codebook gain G, the pitch lag M and the $\{\beta_i\}$ by exhaustively searching for the weighted MMSE (minimum mean square error) between the original speech and the perceptually weighted reconstructed signal; (3) Optimizing the M and $\{\beta_i\}$ by sequentially searching for the MMSE or MSPE (modified minimum squared prediction error) (Kleijn et al., 1988).

The third procedure above is often employed in practice and is termed an analysis-by-synthesis search, but can also be viewed as adding a pitch component from an adaptive codebook. For this sequential analysis-by-synthesis search procedure, we assume that the input of the pitch filter is zero. The output of the pitch filter depends on the output of the previous subframe. We find the optimum M and $\{\beta_i\}$ first. Then, we search for the optimum excitation codewords.

There is a local (recursive) pitch synthesis filter in a CELP coder. The transfer function of the pitch synthesis filter can be expressed as

$$H(z) = \frac{1}{1 - P(z)}, \tag{27}$$

where

$$P(z) = \sum_{i=-1}^{1} \beta_i z^{-M+i}, \tag{28}$$

The input of the pitch synthesis filter, $\hat{e}(z)$ is the codeword from the excitation codebook. We can decompose the excitation codeword into two components: an ideal prediction residual (that would be obtained at the analysis stage, shown in Fig. 1) $e(z)$, and a quantization noise output $q_n(z)$.

$$\hat{e}(z) = e(z) + q_n(z), \tag{29}$$

where

$$e(z) = x(z)(1 - P(z)). \tag{30}$$

The output of the pitch synthesis filter, $\hat{x}(z)$, is

$$\hat{x}(z) = \frac{\hat{e}(z)}{1 - P(z)} = x(z) + \frac{q_n(z)}{(1 - P(z))}. \tag{31}$$

For the first term of (31), the pitch prediction residual, stability is not a problem because of pole/zero cancellation in the analysis and synthesis stages. However, the quantization noise passes through only the pitch synthesis filter. If the pitch filter is not stable, this component leads to an increasing pitch filter output. For simplicity, we suppose the quantization noise to be an additive white noise. An unstable filter can result in a large increase of the output noise. In the sequential search procedure, the pitch filter parameters are chosen before the contribution of the stochastic codebook in CELP is considered. The stochastic contribution can drive the output of an unstable pitch filter to large values. This causes distortion of the reconstructed signal. The enhanced quantization noise can be further augmented in the following subframes, because the adaptive codebook is updated with the accumulated noise of an unstable pitch filter. The average SNR of a CELP coder for an adaptive codebook procedure with a conventional (unquantized pitch coefficient) 3T3DF pitch filter for one test sentence can fall to 3.89 dB, comparing to 9.0 dB for a 1T1DF filter. The waveform of the reconstructed speech with an adaptive codebook for a 3T3DF in several subframes of an unstable pitch filter is compared with the original speech signal (Fig. 15(a)) and is shown in Fig. 15(b). Fig. 15(c) gives the reconstructed waveform with a stabilized pitch filter $3T3DF_{b1.0}$ under a simple sufficient stable condition, as described in the sequel.

A simplified stability test and four stabilization techniques have been proposed to efficiently tame an unstable pitch filter in (Ramachandran and Kabal, 1987). The simple sufficient stability conditions are

$$\begin{aligned} &|\beta| < 1, && 1T1DF, \\ &|\beta_{-1}| + |\beta_0| + |\beta_{+1}| < 1, && 3T3DF. \end{aligned} \tag{32}$$

Let $a = \beta_{-1} + \beta_{+1}$ and $b = \beta_{-1} - \beta_{+1}$. The sufficient stability conditions for a 3T3DF pitch filter are (Ramachandran and Kabal, 1987)

$$\begin{aligned} &(1)\ \text{if } |a| \geqslant |b|, \quad |\beta_{-1}| + |\beta_0| + |\beta_{+1}| < 1. \\ &(2)\ \text{if } |a| < |b| \text{ and } |\beta_0| + |a| < 1; \quad b^2 \leqslant a \text{ or } b^2\beta_2^2 - (1-b^2)(b^2-a^2) < 0. \end{aligned} \tag{33}$$

The tight sufficient conditions reduce to the simple sufficient conditions (32) for both 3T1DF and 3T2DF filters, since $b = 0$ and $|a| > 0$. The 3T1DF pitch filter has a better stability performance than a conventional 3T3DF filter, since we constrain the side prediction coefficients $\beta_{-1} = \beta_{+1}$ to be a small proportion of the center coefficient β_0. It is easier for the 3T1DF filter to meet the sufficient condition for the simplest stability test in (32),

$$|\beta_0| < \frac{1}{1 + 2|\alpha|}. \tag{34}$$

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

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

Since each of $|\gamma|$ and $|\beta_0|$ is possibly larger than 1, the chance that the filter violates the sufficient condition is higher than that for a 3T1DF filter.

A simple stabilization method by scaling the coefficients is used to stabilize the pitch synthesis filter. We scale down the pitch coefficients by multiplying a factor c,

$$c = \frac{V_{\text{th}}}{(|\beta_{-1}| + |\beta_0| + |\beta_{+1}|)}, \quad \text{if } (|\beta_{-1}| + |\beta_0| + |\beta_{+1}|) > V_{\text{th}}. \tag{35}$$

The threshold V_{th} is an experimentally determined threshold. With $V_{\text{th}} = \infty$, no stabilization is used. With $V_{\text{th}} = 1$, a strict stability condition is imposed.

The pseudo-multi-tap pitch filters, 3T1DF and 3T2DF pitch filters were incorporated into an FS1016 4.8 kbit/s CELP coder (Campbell et al., 1990). We employ three performance measures: the average

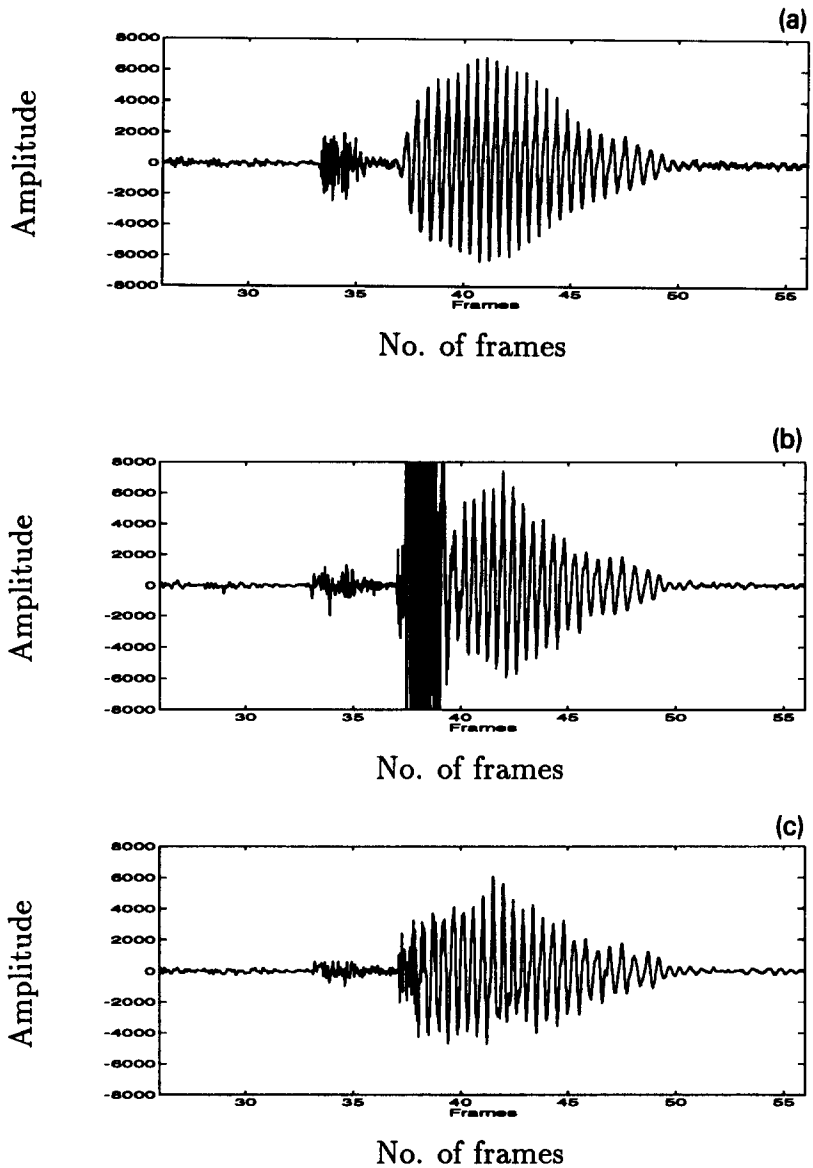

Fig. 15. (a) Original speech waveform. (b) Reconstructed waveforms with an unstable 3T3DF pitch filter. (c) Reconstructed waveforms with a stabilized $3T3DF_{b1.0}$ pitch filter.

SNR, signal-to-noise ratio, the SEGSNR, segmental signal-to-noise ratio (average of log SNRs evaluated for 16 ms) segments and the SFG, the synthesis-filter-gain. We define the SFG as the ratio of the energy of the original speech signal and the energy of the error between the original speech signal 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 3 shows these performance measures for two male and two female test sentences. The test sentences in the experiments are as follows: *Add the sum to the product of these three; Oak is strong and*

Table 3
SNR (dB) comparisons for different pitch synthesis filters in a CELP speech coder

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

also gives shade. Each sentence lasts about three seconds. They were recorded with a 20 kHz sampling rate, 15-bit A/D with Rockland filters set for a cutoff of 5.5 kHz (1 dB down at 5 kHz, 40 dB down at 10 kHz). The files were obtained by digitally filtering the 20 kHz data and changing the sampling rate to 8 kHz. For comparison, the performance for a conventional one-tap filter (1T1DF) and a conventional 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 stability threshold V_{th} is set to be 1.0, 1.10, 1.15 and 2.0 for the comparisons. The threshold V_{th} is denoted in the subscript of the type of the pitch filter. For example, $1T1DF_{b1.0}$ and $3T1DF_{b1.15}$ employ thresholds of 1.0 and 1.15, respectively. The $3T1DF_{b2.0}$ filter obtains an SNR increase of 1.16 dB, compared to a $1T1DF_{b1.0}$ filter. The $3T1DF_{b\infty}$ and $1T1DF_{b\infty}$ configurations use $V_{\text{th}} = \infty$. This means that the pitch filter is not stabilized.

The results show that the stabilization can actually increase the performance for a particular pitch filter configuration. Moreover, a relaxed stability constraint is better than a strict stability constraint. The reason is that the increasing pitch pulse amplitudes are needed to model a 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 a $3T1DF_{b2.0}$ filter and a conventional $3T3DF_{b2.0}$ filter is only 0.32 dB. The performance of a $3T1DF_{b2.0}$ filter is close to a $3T3DF_{b2.0}$ filter.

In addition to objective SNR measurements, we have ranked the subjective quality using informal listening tests. The $3T3DF_{b2.0}$ configuration gives the best quality. The $3T1DF_{b2.0}$ filter offers more natural speech than a $1T1DF_{b2.0}$ filter. The $3T3DF_{b\infty}$ configuration is the worst, because of the stability problems. There are annoying pops, clicks and a more dominant background noise for this case. The $3T2DF_{b\infty}$ filter has the same problem as $3T3DF_{b\infty}$. Although $3T1DF_{b\infty}$ and $1T1DF_{b\infty}$ both have the stability problems, the resultant speech for the latter is not as contaminated as in the other cases.

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 a 3T1DF pitch filter with fractional pitch lags and pitch quantizer, defined in the FS1016 CELP coder. Note that these fractional pitch lags are not uniformly spaced – small lags have higher resolution than large lags. The results show that the SNR and SEGSNR are higher than that of the standard FS1016 coder by 0.45 dB and 0.1 dB. An informal listening test shows that the improved CELP coder with 3T1DF pitch filter is slightly better than the original FS1016 CELP coder.

8. Summary and discussion

We have presented and analyzed two pseudo-multi-tap pitch prediction filter configurations, 3T2DF and 3T1DF. The pseudo-multi-tap pitch filters can be viewed as a shape/gain decomposition, with the 3T1DF filter having only one shape and the switching 3T1DF filter having two shapes. This then reduces the multi-tap coding problem to a scalar quantization of the gain value. The prediction gain of pseudo-three-tap pitch prediction filter is higher than that of a one-tap pitch prediction filter. The frequency response is more desirable than a conventional one-tap and three-tap pitch synthesis filter because of the symmetrical and small side prediction coefficients. The pseudo-multi-tap pitch filter can also be used in the synthesis stage of a speech coder, with the optimal lag and coefficients determined using an analysis-by-synthesis approach. Stabilization, using a relaxed stability criterion, is applied by scaling the pitch filter coefficients. Coefficient scaling based on a relaxed sufficient constraint allows for weakly unstable pitch synthesis filters, which can track fast changing segments during an unvoiced to voiced onset. The pseudo-multi-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 performance of a US federal standard FS1016 4.8 kbit/s CELP coder with a pseudo-multi-tap pitch filter is better than that with a conventional one-tap pitch filter.

References

J.P. Campbell, T.E. Tremain and V.C. Welch (1990), "The proposed federal standard 1016 4800 bps voice coder: CELP", *Speech Technology*, pp. 58–64.

R. Crochiere and L. Rabiner (1983), *Multirate Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ.

V. Iyengar and P. Kabal (1991), "A low delay 16 kbits/s speech coder", *IEEE Trans. Signal Process.*, Vol. 39, pp. 1049–1057.

W.B. Kleijn, D.J. Krasinski and R.H. Ketchum (1988), "Improved speech quality and efficient vector quantization in SELP", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., New York, NY, 11–14 April 1988*, pp. 155–158.

P. Kroon and B.S. Atal (1991), "Pitch predictors with high temporal resolution", *IEEE Trans. Signal Process.*, Vol. 39, pp. 733–735.

R. Ramachandran and P. Kabal (1987), "Stability and performance analysis of pitch filters in speech coders", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35, pp. 937–946.

R. Ramachandran and P. Kabal (1989), "Pitch prediction filters in speech coding", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 37, pp. 467–478.

M. Schroeder and B. Atal (1985), "Code-excited linear prediction (CELP): High quality speech at very low bit rates", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Tampa, FL, 26–29 March 1985*, pp. 937–940.