A TECHNIQUE OF MULTI-TAP LONG TERM PREDICTOR (LTP) FILTER USING SUB-SAMPLE RESOLUTION DELAY

Mark A. Jasiuk, Tenkasi Ramabadran, Udar Mittal, James P. Ashley, Michael J. McLaughlin

Speech Processing Research Lab, Motorola Labs

ABSTRACT

The method of a 1st order long-term predictor (LTP) filter, using a sub-sample resolution delay, is extended to a multi-tap LTP filter, or, equivalently, the conventional integer-sample resolution multi-tap LTP filter is extended to use sub-sample resolution delay. Defining the delay with sub-sample resolution enables this novel multi-tap LTP filter to explicitly model delay values that have a fractional component. The filter coefficients, largely freed from implicitly modeling the effect of delays that have a fractional component, seek to maximize the prediction gain of the LTP filter by modeling the frequency dependent gain. This is in contrast to a conventional multi-tap LTP filter, which applies a single model to tackle the dual tasks of representing the non-integer valued delays and the frequency dependent gain. Experimental results are presented for narrowband and wideband speech. This technique is part of the 3GPP2 Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB) Rate Set 1 Standard.

1. INTRODUCTION

Digital speech coders based on the Analysis-by-Synthesis (A-by-S) paradigm typically employ long-term (pitch) and short-term (formant) predictors that model the characteristics of an input speech signal and that are incorporated into a set of time-varying linear filters. The filter parameters and the filter excitation are quantized, instead of the individual input speech samples.

Code Excited Linear Prediction (CELP)[1] is one example of such a coder. In a CELP coder, an excitation signal for the filters is chosen from a codebook of codevectors. While this paper addresses the long-term predictor (LTP) component of an A-by-S system using CELP as an example, the technique being presented is applicable to any system which uses an LTP.

In Section 2, an overview of prior-art LTP filter configurations is presented. In Section 3, the new LTP filter configuration is introduced with Section 4 containing experimental results comparing the LTP filter of Section 3 to selected prior-art LTP filter configurations. Section 5 provides a summary.

2. LTP FILTERS- PRIOR ART

The synthetic combined excitation for a CELP coder is typically expressed as

$$ex(n) = \gamma \tilde{c}_I(n) + \sum_{i=-K_1}^{K_2} \beta_i ex(n-L+i), 0 \le n < N \qquad (1)$$

where $K_1 \ge 0, K_2 \ge 0$. $\tilde{c}_I(n)$ is a codevector, or excitation vector, selected from a codebook, I is an index specifying the selected codevector, γ is the gain for scaling the selected codevector, $ex(n-L+i)$ is a synthetic combined excitation signal delayed by L integer resolution samples relative to the $(n+i)^{th}$ sample of the current subframe (for voiced speech L is typically related to the pitch period), β_i's are the long term predictor (LTP) filter coefficients, and N is the number of samples in the subframe. When $n - L + i < 0$, $ex(n-L+i)$ contains the history of past synthetic excitation, constructed prior to the current subframe. The LTP filter transfer function is given by

$$P(z) = \frac{1}{1 - \sum_{i=-K_1}^{K_2} \beta_i z^{-L+i}}, K_1 \ge 0, K_2 \ge 0 \qquad (2)$$

where the LTP filter order is $K = (K_1+K_2+1)$[2][3].

The task of a typical CELP speech coder is to select the parameters specifying the synthetic excitation, that is, the parameters $L, \beta_i's, I,$ and γ, given $ex(n)$ for n < 0 and the determined coefficients of short-term Linear Predictor (LP) filter, so that when the synthetic excitation sequence $ex(n)$ for $0 \le n < N$ is filtered through the LP filter, the resulting synthesized speech signal $\hat{s}(n)$ most closely approximates, according to a distortion criterion employed, the input speech signal s(n) to be coded for that subframe.

When the LTP filter order $K > 1$, the LTP filter as defined in (2) is a multi-tap filter. A conventional integer-sample resolution delay multi-tap LTP filter seeks to predict a given sample as a weighted sum of K, usually

adjacent, delayed samples, where the delay is confined to a range of expected pitch period values (typically between 20 and 147 samples at 8 kHz signal sampling rate). A multi-tap LTP filter requires quantization of the K unique β_i coefficients, in addition to L.

If $K = 1$, a 1st order LTP filter results, requiring quantization of only a single β_0 coefficient and L. However, a 1st order LTP filter, using integer-sample resolution delay L, does not have the ability to model a non-integer delay value, other than rounding it to the nearest integer or selecting an integer multiple of a non-integral delay. Neither does it provide any frequency dependent gain (also called spectral shaping). Nevertheless, 1st order LTP filter implementations have been commonly used, because only two parameters - L and β_0 - need to be quantized, a consideration for many low-bit rate speech coder implementations.

The introduction of the 1st order LTP filter, using a sub-sample resolution delay, significantly advanced the state-of-the-art of LTP filter design [4][5][6]. Using this technique, the delay value L is explicitly represented with sub-sample resolution, redefined here as $\hat{L}$. Samples delayed by $\hat{L}$ may be obtained by using an interpolation filter. Such a 1st order LTP filter is able to provide predicted samples with sub-sample resolution, but lacks the ability to provide spectral shaping. The LTP filter transfer function for this filter is given by

$$P(z) = \frac{1}{1-\beta_0 z^{-\hat{L}}}, 0 \le n < N \qquad (3)$$

with the corresponding difference equation given by:

$$ex(n) = \gamma \tilde{c}_I(n) + \beta_0 ex(n-\hat{L}), 0 \le n < N \qquad (4)$$

Implicit in equations (3) and (4) is the use of an interpolation filter to compute samples pointed to by the sub-sample resolution delay $\hat{L}$.

Note that in describing the LTP filter, a generalized form of the LTP filter transfer function has been given. *ex(n)* for values of $n < 0$ contains the LTP filter state. For values of $\hat{L}$ which necessitate access to samples indexed by $n \ge 0$ when evaluating *ex(n)* in eqn. (4) (or in eqn. (1)), a simplified and non-equivalent form for the LTP filter is often used, called a virtual codebook or an adaptive codebook (ACB)[7]. Extending the ACB technique, which was described in [7] in the context of integer valued L, to a sub-sample resolution $\hat{L}$, eqn. (4) is redefined as follows by eqns. (5a)-(5c):

$$ex(n) = ex(n-\hat{L}), 0 \le n < N \qquad (5a)$$

$$c_0(n) = ex(n), 0 \le n < N \qquad (5b)$$

$$ex(n) = \beta_0 c_0(n) + \gamma \tilde{c}_I(n), 0 \le n < N \qquad (5c)$$

In eqn. (5a), *ex(n)* is extended for $0 \le n < N$, with the extended vector becoming the ACB vector, $c_0(n)$, of eqn. (5b).

Considering two of the LTP filter configurations previously discussed; i.e., an integer-sample resolution delay multi-tap LTP filter and a 1st order sub-sample resolution delay LTP filter, the following observations may be made:

The conventional multi-tap predictor performs two tasks simultaneously: spectral shaping and implicit modeling of a non-integer delay through generating a predicted sample as a weighted sum of K delayed samples [2][3]. In the conventional multi-tap LTP filter using integer resolution delays, these two tasks are inextricably tied together.

The 1st order sub-sample resolution LTP filter, on the other hand, can explicitly use a fractional part of the delay to select a phase of an interpolating filter of a high order. This method, where the sub-sample resolution delay is explicitly defined and used, provides a very efficient way of representing interpolation filter coefficients. Those coefficients do not need to be explicitly quantized and transmitted, but can be inferred from the delay received, where that delay is specified with sub-sample resolution. While such a filter does not have the ability to introduce spectral shaping, for narrowband, voiced (quasi-periodic) speech, it has been found that the effect of defining the delay with sub-sample resolution is more important than the ability to introduce spectral shaping [5]. These are some of the reasons why a 1st order LTP filter, with a sub-sample resolution delay, is widely used in numerous speech codec standards.

While a sub-sample resolution 1st order LTP filter is a very efficient model for representing non-integer delays, it may be desirable, in addition, to provide a mechanism for incorporating spectral shaping. The speech signal harmonic structure tends to weaken at higher frequencies. This effect becomes more pronounced for wideband speech coding systems, characterized by an increased signal bandwidth of 8 kHz relative to 4 kHz of narrow-band signals. One method of adding spectral shaping is described in [8]. This approach provides two spectral shaping filters to select from and requires that the ACB vector be explicitly filtered by the spectral shaping filter being evaluated. The filtered version of the ACB vector is then used to generate a distortion metric, which is evaluated to select a spectral shaping filter to use in conjunction with the LTP filter parameters. If a large set of spectral shaping filters is provided to select from, this may result in appreciable increase in complexity due to the filtering operations. Also, the information related to the selected filter, such as an index m, needs to be quantized and conveyed from the encoder to the decoder.

3. MULTI-TAP LTP FILTER USING SUB-SAMPLE RESOLUTION DELAY

In this section a multi-tap LTP filter, using a sub-sample resolution delay $\hat{L}$, is presented. The generalized transfer function of the new LTP filter is:

$$\frac{1}{1-\sum_{i=-K_1}^{K_2}\beta_i z^{-\hat{L}+i}}, K_1 \geq 0, K_2 \geq 0, K = 1+K_1+K_2 > 1 \quad (6)$$

Selecting $K > 1$, results in a K^{th} order multi-tap LTP filter. The coefficients, β_i's, may be computed or selected to maximize the prediction gain of the LTP. In addition to implicitly fine tuning the sub-sample resolution delay $\hat{L}$, the β_i's coefficients embody the spectral shaping characteristic; that is, there need not be a dedicated set of spectral shaping filters to select from, with the filter selection decision then quantized and conveyed from the encoder to the decoder. Moreover, no explicit filtering needs to be done to compute the distortion metric corresponding to a β_i vector being evaluated, as was shown in [9] for a conventional multi-tap LTP filter.

If desired, the LTP filter coefficients may be entirely prevented from implicitly fine tuning the sub-sample resolution delay $\hat{L}$, by requiring the taps of the LTP filter to be symmetric; i.e., $\beta_{-i} = \beta_i$ for $-K_1 \leq i \leq K_2$ where $K_1 = K_2$ and K is odd. Such a configuration may be advantageous for quantization efficiency and to reduce computational complexity.

The CELP generalized difference equation, corresponding to eqn. (6), for creating the combined synthetic excitation *ex(n)*, is:

$$ex(n) = \gamma \tilde{c}_I(n) + \sum_{i=-K_1}^{K_2}\beta_i ex(n-\hat{L}+i), \quad 0 \leq n < N \quad (7)$$

Eqn. (7) may be modified to use ACB implementation, as follows:

$$ex(n) = ex(n-\hat{L}), \quad 0 \leq n < N+K_2 \quad (8a)$$

Using samples of *ex(n)* generated in eqn. (8a), a new signal $c_i(n)$ is defined:

$$c_i(n) = ex(n+i), 0 \leq n < N, -K_1 \leq i \leq K_2 \quad (8b)$$

The combined synthetic subframe excitation, embodying the ACB implementation, may now be expressed, using the results from eqns. (8a)-(8b), as:

$$ex(n) = \gamma \tilde{c}_I(n) + \sum_{i=-K_1}^{K_2}\beta_i c_i(n), 0 \leq n < N \quad (8c)$$

Let $p(n)$ be the perceptually weighted target vector, with the zero input response of the perceptually weighed synthesis filter subtracted out, and $\tilde{c}_I'(n)$ and $c_i'(n)$ be the filtered (by the zero state perceptually weighted synthesis filter) versions of $\tilde{c}_I(n)$ and $c_i(n)$ respectively. The perceptually weighted subframe error energy, E, is:

$$E = \sum_{n=0}^{N-1}\left[p(n) - \gamma \tilde{c}'_I(n) - \sum_{i=-K_1}^{K_2}\beta_i c'_i(n)\right]^2 \quad (9)$$

The task of the CELP speech coder is to select the parameters- $\hat{L}$, β_i's, I, and γ, so that E is minimized, subject to the excitation parameter search constraints employed. For example, once the excitation vectors are selected, the optimal set of coder gains may be jointly computed, by solving a system of K+1 simultaneous equations:

$$\frac{\partial E}{\partial \beta_i} = 0, \quad \text{for} -K_1 \leq i \leq K_2, \text{ and } \frac{\partial E}{\partial \gamma} = 0 \quad (10)$$

Alternately, if the coder gains β_i and γ are quantized, using scalar quantization, vector quantization, or some combination of the two, minimization of E may be used as the criterion for selecting quantized versions of those parameters. Note that eqns. (9)-(10) may be restated to use the pre-computed correlation terms among the vectors $p(n)$, $\tilde{c}_I'(n)$, and $c_i'(n)$ to reduce computational complexity, employing the approach of [9]. A technique for optimally selecting $\tilde{c}_I'(n)$, given the K selected $c_i'(n)$ vectors, is presented in [10].

4. EXPERIMENTAL RESULTS

To evaluate the performance of the new LTP filter configuration over narrowband and wideband sampled speech, average LTP prediction gain was computed using LP residual as input, following an approach outlined in [5]. An ACB implementation of the LTP was used to generate the prediction gain data.

A 221 second database, containing 4 male and 4 female speakers, sampled at 16 kHz and digitally band-pass filtered to 50-7000 Hz, was used to generate the wideband results. That database was in addition digitally band-pass filtered to 100-3600 Hz, and decimated by a factor of two, for use in the narrowband experiments. LP coefficients were computed for every 5 ms frame, using a 20 ms rectangular window and covariance analysis, and the corresponding LP residual signal was generated. 10^{th} and 20^{th} order LP analysis was selected for the narrowband and wideband speech respectively. The input speech was analyzed by a speech voicing classifier to exclude silence and strongly unvoiced frames and to ensure that the same subset of 5 ms frames contributed to the prediction gain computation for each condition.

Table 1 illustrates the average prediction gain results. K' is the number of unique filter coefficients and D specifies the oversampling factor. In the case where the filter coefficients are independent, the filter order K is

equal to K'. Where the filter coefficients are constrained to be symmetric, the filter order is $K = 2(K'-1)+1$ and the corresponding prediction gain values are shown inside the parentheses.

Table 1. Average LTP Prediction Gain Values

K'	D	Average Prediction Gain (dB)			
		Narrowband input		Wideband input	
		female	male	female	male
1	1	5.9	5.0	2.6	2.7
1	2	7.2	6.3	3.1	3.3
1	4	8.0	6.9	3.3	3.5
1	8	8.3	7.1	3.4	3.5
2	1	7.1 (6.3)	6.3 (5.4)	3.4 (3.2)	3.5 (3.1)
3	1	7.9 (6.4)	7.2 (5.7)	3.8 (3.3)	3.9 (3.3)
2	2	7.8 (7.4)	7.0 (6.6)	3.7 (3.5)	3.7 (3.6)
3	2	8.4 (7.6)	7.7 (6.9)	4.0 (3.7)	4.0 (3.8)
2	4	8.4 (8.2)	7.5 (7.3)	3.8 (3.7)	3.8 (3.8)
3	4	8.9 (8.4)	8.0 (7.6)	4.1 (3.9)	4.1 (4.0)
2	8	8.7 (8.5)	7.6 (7.5)	3.8 (3.8)	3.8 (3.8)
3	8	9.1 (8.8)	8.1 (7.8)	4.1 (3.9)	4.1 (4.0)

Comparing the wideband and narrowband results, the lower wideband prediction gains can be attributed to the weaker harmonic structure above 4 kHz. The first 4 rows illustrate the increase in prediction gain due to progressively higher oversampling factors. Rows 5 and 6, illustrate the prediction gains of conventional integer resolution delay multi-tap predictors. Rows 7 through 12 demonstrate the prediction gains of the proposed multi-tap LTP filter using sub-sample resolutions. In general, for a given D, increasing K' (and similarly, for a given K', increasing D) results in a monotonic increase in prediction gain. Forcing the coefficients to be symmetric, results in a slight drop in prediction gain, as expected.

5. CONCLUSIONS

A novel formulation of an LTP filter was presented, which extends the technique of a sub-sample resolution delay 1st order LTP predictor to a multi-tap LTP filter, providing a flexible framework for explicit modeling of sub-sample resolution delay $\hat{L}$, thus freeing the β_i coefficients to mainly model the frequency dependent gain. This gives the algorithm designer an opportunity to trade off the selection of D vs. K when optimizing the LTP filter configuration for a given system. The integration of this technique into a practical speech compression system was illustrated in the context of a generic CELP coder example. The performance of the new LTP filter structure was demonstrated, for narrowband and wideband inputs, by computing average prediction gain values for a number of LTP filter configurations. This technique is part of the 3GPP2 VMR-WB Rate Set 1 Speech Codec Standard [8].

6. REFERENCES

[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. Acoustics, Speech, and Signal Processing*, VOL. 3, pp. 937-940, 1985.

[2] B. S. Atal, "Predictive Coding of Speech at Low Bit Rates," *IEEE Transactions on Communications*, VOL. COM-30, NO. 4, pp. 600-614, April 1982.

[3] R. P. Ramachandran and P. Kabal, "Pitch Prediction Filters in Speech Coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, VOL. 37, NO. 4, pp. 467-478, April 1989.

[4] I. A. Gerson and M. A. Jasiuk, "Digital Speech Coder Having Improved Sub-sample Resolution Long-Term Predictor," US Patent No. 5,359,696.

[5] P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, VOL. 2, pp. 661-664, 1990.

[6] J. S. Marques, I. M. Trancoso, J. M. Tribolet, L. B. Almeida, "Improved Pitch Prediction with Fractional Delays in CELP Coding," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, VOL. 2, pp. 665-668,1990.

[7] W. B. Kleijn, D. J Krasinski, and R. H. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, VOL. 1, pp. 155-158, 1988.

[8] "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Option 62 and 63 for Spread Spectrum Systems," Document 3GPP2 C.P0052-A, Version 0.3, December 10, 2004.

[9] J.-H. Chen, "Toll-Quality 16 kb/s CELP Speech Coding with Very Low Complexity," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, VOL. 2, pp. 9-12, 1995.

[10] U. Mittal, J. P. Ashley, E. M. Cruz-Zeno, and M. A. Jasiuk, "Joint Optimization of Excitation Parameters in Analysis-by-Synthesis Speech Coders Having Multi-tap Long Term Predictor," *to be presented at ICASSP 2005.*