S12.7

IMPROVED PITCH PREDICTION WITH FRACTIONAL DELAYS IN CELP CODING

J. S. Marques, I. M. Trancoso, J. M. Tribolet and L. B. Almeida

INESC / IST, R. Alves Redol, 9, P-1000, Lisbon, Portugal

ABSTRACT

A broad class of coders uses long-term prediction to exploit the quasi-periodic structure of voiced speech. The pitch period varies smoothly and hence should be modeled by a real number. However, in coding applications, the predictor delay is usually restricted to integer multiples of the sampling interval. At low bit rates, this restriction degrades the performance of the coder, especially for high pitched sounds. This paper discusses an alternative scheme for long-term prediction in CELP coding using fractional delay prediction. This technique enables a more accurate representation of voiced speech and achieves an improvement of synthetic quality for female speakers.

INTRODUCTION

Nowadays, one of the most common synthesis models uses of an almost "white" signal to excite two linear filters: one that takes into account the periodic structure of voiced speech and another which imposes a non-flat spectral envelope [3]. Coders based on this kind of model started with adaptive predictive coding (APC) [4] and include such structures as multipulse (MP), regular pulse excitation (RPE) and code excited linear predictive (CELP) coding [5], for instance. At high bit rates, enough bits are assigned to the excitation signal in these coders, to enable them to reconstruct the harmonic structure that the long term predictor fails to accurately model. At low bit rates, however, the excitation does not have so much variability and the synthetic speech quality is much more dependent on the performance of the long term predictor. The pitch period varies smoothly and hence should be modeled by a real number. However, in coding applications, it is usually restricted to integer multiples of the sampling interval. This restriction has more pronounced effects on high pitched sounds, resulting in the partial destruction of the harmonic structure, especially in the high frequency regions [2].

Frequently, two- and three-tap predictors have been adopted in order to simultaneously provide an "interpolated" value for the non-integer delay and a frequency dependent gain factor. However, this has obvious disadvantages in terms of bit rate and may pose stability problems. In addition, the computation of the multiple taps is usually performed in a suboptimal way, by first determining the optimal delay assuming a single tap predictor, and then evaluating the scale factors for neighbouring delays. An alternative solution consists of adopting fractional values for the pitch period. This concept has been proposed in [1] and will be extensively studied here.

There are several schemes for integrating pitch prediction in CELP coders. Traditionally, the long-term predictor coefficients were computed in an open-loop configuration, i.e., on the basis of the original speech signal residual. Recently, however, a significant step towards increasing robustness has been done, by extending the analysis-by-synthesis procedures used in the determination of the excitation signal in CELP and multipulse, to the computation of the long-term predictor and taps [6].

The objective of this paper is the study of the pitch-related issues in CELP-based coders. The emphasis will be on the use of fractional delay predictors, as compared with classical integer-delay prediction, but we shall also discuss the use of single and multiple-tap predictors, open and closed-loop configurations, etc. This is an important subject, given the ongoing standardization work involving CELP coders in Europe and U.S.A. and the fact that a significant part of the distortion in this kind of coder is due to poor modeling of the harmonic structure.

PREDICTION WITH FRACTIONAL DELAYS

In voiced sounds, the vocal tract is excited by the vibration of the vocal folds, periodically interrupting the airflow at the glottis. Since the movements of the vocal tract articulators (tongue, jaw, etc.) are slow compared to the pitch period, the speech signal produced by the vibration of the vocal folds is quasi-periodic and therefore highly predictable. The most basic long-term predictor is the one tap filter given by

$$s_p(n) = G\, s(n-N) \qquad (1)$$

where s(n) is the input signal, $s_p(n)$ is the predicted signal, N is the integer delay value, and G is a gain, usually computed by least squares techniques. Since delays are restricted to integer values, this predictor is not able to cope with arbitrary pitch intervals without replacing, in some way, the optimal non-integer delay by an integer value, which degrades the performance of the predictor in terms of SNR.

When trying to generalize equation (1) to fractional delay values, the plain substitution of the delay N by a real number D is not possible, as the discrete-time signal s(n) is not defined for non-integer values of the argument. To overcome this difficulty, s(n) is replaced in equation (1) by an interpolated signal $s_c(t)$ which coincides with s(n) at integer argument values. The expression of

CH2847-2/90/0000-0665 $1.00 © 1990 IEEE

the one-tap predictor with non-integer delays is therefore given by

$$s_p(n) = G\, s_c(n-D) \qquad (2)$$

Assuming that the continuous-time speech signal is sampled according to the Nyquist criterion, it can be recovered at any time instant by filtering the discrete-time signal s(n) through an ideal non-causal low-pass filter:

$$s_c(t) = \sum_{k=-\infty}^{\infty} s(k)\, Sa(\pi(t-k)) \qquad (3)$$

where Sa(x) is the sampling function sin(x)/x.

By substituting (3) into (2), one obtains

$$s_p(n) = G \sum_{k=-\infty}^{\infty} s(k)\, Sa(\pi(n-D-k)) \qquad (4)$$

In practical applications, the summation is restricted to a finite number of terms centered around the integer nearest to n-D (details about the truncation can be found in [1]).

FRACTIONAL DELAY ESTIMATION

One of the most common methods for delay estimation is based on the minimization of a quadratic cost functional, due to its mathematical tractability. Let J(D) be the cost functional given by

$$J(D) = \sum_{n=n_b}^{n_b+L-1} \left[s(n) - G(D)\, s_c(n-D) \right]^2 \qquad (5)$$

where n_b is the beginning of the frame and L is the length of the optimization interval and G(D) is the optimal predictor gain, which is dependent on the delay D. Minimization in order to G(D) yields the following expression for the optimum gain:

$$G(D) = \frac{\sum_{n=n_b}^{n_b+L-1} s(n)\, s_c(n-D)}{\sum_{n=n_b}^{n_b+L-1} s_c(n-D)^2} \qquad (6)$$

Substituting this gain in (5) and discarding the fixed term, the minimization of the quadratic cost functional becomes equivalent to the maximization of the functional C(D) given by

$$C(D) = \frac{\left(\sum_{n=n_b}^{n_b+L-1} s(n)\, s_c(n-D) \right)^2}{\sum_{n=n_b}^{n_b+L-1} s_c(n-D)^2} \qquad (7)$$

Hence, the choice of the best delay value consists of the determination of the global maximum of C(D). When the delay is restricted to integer values, the maximum is easily obtained by computing this function for a grid of integer delays, and then choosing the largest value. However, with real-valued delays, there is no a priori knowledge of the bandwidth of the continuous-time function C(D), and therefore sampling the delay with integer values may not be enough to characterize the function. Furthermore, even if C(D) were sampled according to the Nyquist criterion, the maximum of the sampled signal could be located near local maxima of the continuous-time signal instead of being in the neighbourhood of the global maximum. This is illustrated in figure 1 which shows the LPC residue of a voiced speech segment and the function C(D) used to evaluate the optimal delay in the interval from 20 to 27.5 ms. The speech signal was sampled at 8 KHz and the function C(D) shown in figure 1b was evaluated with 8 fractional delays per sampling interval. Figure 1c displays 8 down-sampled versions of figure 1b with only one delay per sampling interval (125 μs). They represent the function C(D) computed at integer delays in the left side version, and at fractional delays in the remaining ones. The interval between consecutive delays is constant but each version starts at different delay values

$$D^j = \frac{1}{64} j \quad \text{ms} \qquad j=0,1,2,\ldots,7 \qquad (8)$$

Depending on the delay grid used to evaluate the function C(D), the maximum will correspond to one pitch period (1st function), two pitch periods (3rd function) or even 3 pitch periods (5th function). This mechanism may therefore be responsible for the occurrence of many delay jumps in long-delay prediction of speech. Hence, the use of a very fine delay grid is instrumental to avoid missing the global maximum.

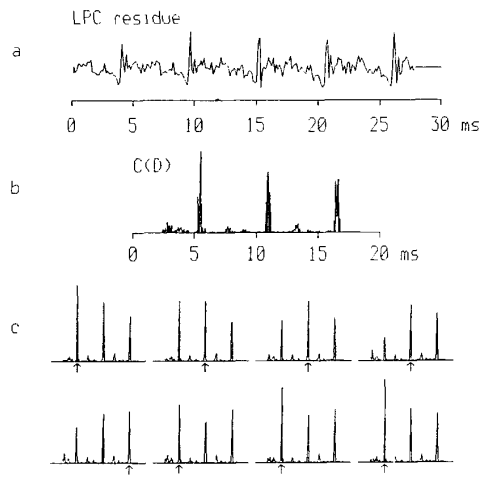

Fig. 1 - Delay Estimation: a) LPC residual;
b) function C(D); c) down-sampled versions of C(D)

In the above figure, an equispaced delay grid was adopted. However, since we are attempting to improve the performance of

the long-delay predictor especially in high-pitched voices, we are naturally led to try a grid which is denser for short delays than for longer ones, such as a logarithmic grid, for instance. Figure 2 shows an example where we compare the delay evolution in a female speech utterance using integer delays and uniformly or logarithmically quantized fractional delays. In (a), the optimal delay is evaluated using a grid of 128 integer values ranging from 20 to 147 sampling intervals (2.5 to 18.5 ms). We can distinguish the pitch contour and a large number of jumps to multiples of the pitch period. Notice that some of the jumps may occur, not because the speech signal is more correlated with the same signal two or three pitch periods before, but because the integer delays are not close enough to the global maximum of C(D). In (b), the same delay range is filled with 256 fractional values uniformly distributed in the same interval, and in (c) the delays are distributed according to a logarithmic law leading to a much denser grid for short delays. In both cases, a significant reduction on the number of jumps can be observed relative to the integer delay grid, achieving a much smoother trajectory. The performance of the last predictor scheme is only slightly better than the previous one.

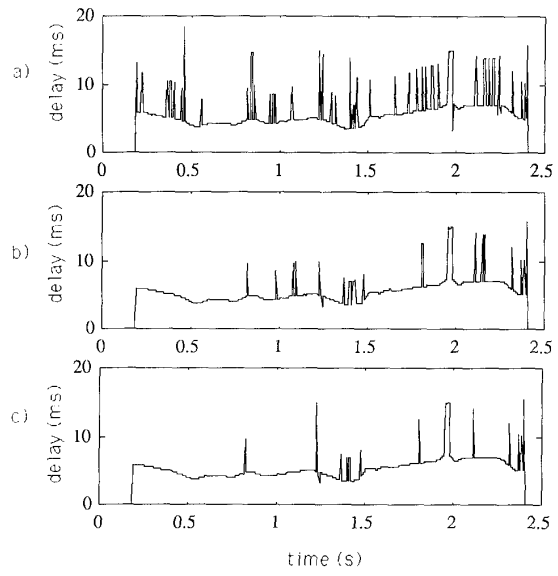

Fig. 2 - Evolution of the optimal delay in long-term prediction a) with integer delays; b) with fractional delays (uniform distribution); c) with fractional delays (logarithmic distribution)

The integration of long-term prediction with fractional delays in a CELP based coder raises two different kinds of problems. In fact, owing to the non-causal filtering operation implicit in the interpolation procedure, the knowledge of the input in the present frame may be necessary, for short delays. In open-loop schemes, the predictor input is the LPC residual which is known in the present frame as well as in the past ones. Therefore, the integration of long-term prediction with fractional delays in open-loop configurations is straightforward. However, the adoption of a closed-loop, analysis-by-synthesis procedure for determining the optimum fractional delay is not so easy. In this case, one does not know the input signal values beyond the beginning of the present frame. To overcome this difficulty we can assume a zero input in the present frame or recursively predict the unknown values using the long-term predictor [1]. In this paper we have adopted the second choice.

The second problem is related to the use of recursiveness in the search for the optimal delay. When using fractional delays, one cannot always find recursive relationships which are usually adopted to decrease the complexity of the delay search [6]. In open-loop configurations one can interpolate the input signal prior to the delay search, provided the delay is uniformly quantized. In this case, one simply applies standard search techniques to the oversampled input signal. In closed-loop configurations, we have not explored recursiveness yet.

EXPERIMENTAL RESULTS

The experimental evaluation of the performance of pitch prediction with fractional delays in CELP coding has been done using the unquantized reference coder described in a companion paper [7]. The frame length was chosen equal to 5 ms, the LPC filter was updated every 20 ms and the long-term predictor every 5 ms. The stochastic excitation was obtained from a 10-bit codebook with Gaussian codewords center-clipped at 1.3. The test data base consisted of 4 phoneme-specific utterances, each spoken by 4 male and 4 female speakers, amounting to a total of 90 s of speech, approximately.

Two basic schemes were compared in this CELP coding framework: long-term prediction with integer delays (with 1 and 2-taps), and fractional delay prediction with both linearly and logarithmically quantized delays. The first type of predictors uses 128 integer delay values ranging from 20 to 147 sampling intervals while the second one uses 256 fractional delays distributed in the same range. Each of the prediction schemes was tested in open- and closed-loop configurations.

The difference between integer and fractional delay prediction is illustrated in figure 3, which shows spectrograms of original and synthetic voiced speech utterances. Figure 3b was obtained with a single tap predictor with integer delays. One can easily notice the increase of inter-harmonic noise in the lower region of the spectrum and the destruction of the harmonic structure in the higher frequencies. Figure 3c corresponds to the use of fractional delays which enable a better representation of the harmonic structure above 2 KHz.

Table I summarizes segmental SNR results. For each predictor, we discriminate the values obtained for male and female speakers and, in each case, we also distinguish the performance of the coder in voiced frames and in whole utterances. One can conclude that fractional delay prediction has always increased SNR results over 1-tap integer delay prediction, performing however slightly worse than 2-tap predictors. The objective improvement is more pronounced for female speakers in voiced frames.

Informal listening tests were carried out to assess the performance of the prediction schemes. In female utterances, there was a clear preference for the results produced with

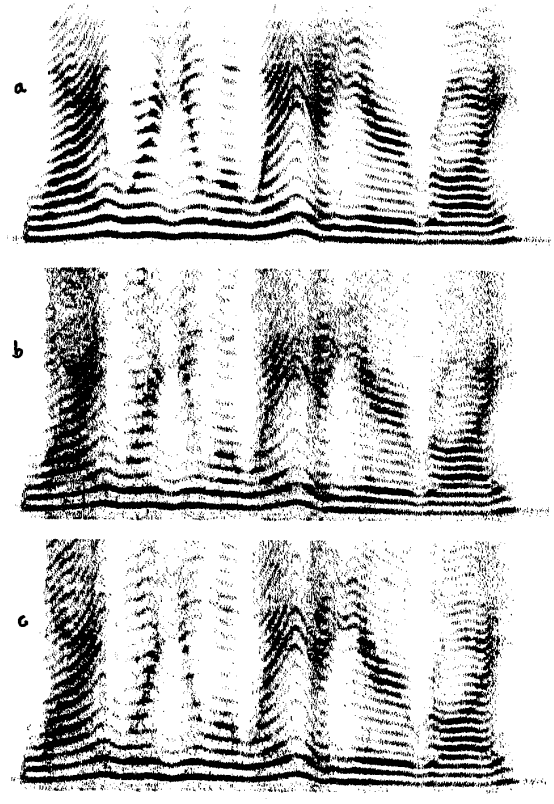

Fig. 3 - Spectrogram of a female utterance: a) original; b) CELP with integer delays; c) CELP with fractional delays

	OPEN-LOOP				CLOSED-LOOP			
	i-1tap	i-2tap	f1-1tap	f2-1tap	i-1tap	i-2tap	f1-1tap	f2-1tap
male speakers	10.2 11.9	10.6 12.5	10.5 12.4	10.4 12.3	11.7 13.2	12.3 14.0	12.1 13.6	12.0 13.5
female speakers	11.8 13.3	12.4 14.1	12.0 13.6	12.2 13.8	13.4 14.7	14.0 15.4	13.7 15.1	13.8 15.2
male and female speakers	11.0 12.6	11.5 13.3	11.3 13.0	11.3 13.0	12.6 13.9	13.2 14.7	12.9 14.4	12.9 14.4

11.5 whole utterances
13.3 voiced frames

Table I - Segmental SNR Results in CELP Coding
Performance of long-delay predictors in open and closed-loop:
i-1tap: integer delays with 1 tap; *i-2-tap:* integer delays with 2 taps;
f1-1tap: fractional delays with 1 tap (linear distribution);
f2-1tap: fractional delays with 1 tap (logarithmic distribution).

fractional delays, over the results of the 1-tap predictor with integer delays, especially in open-loop predictors. When compared with the 2-tap predictor, the single tap predictor with fractional delays was judged to be equivalent in the closed-loop case, but slightly worse in open-loop. These conclusions are also valid in male utterances. However, in this case, the differences between the performance of the predictors become smaller. Concerning the logarithmic vs. uniform quantization of the delay in fractional delay prediction we have observed a small improvement in female utterances by the use of the logarithmic law.

CONCLUSIONS AND FUTURE TRENDS

The use of long-term prediction with non-integer delays achieves an improved accuracy in the representation of voiced speech. The harmonic structure above 2 KHz is emphasized and the synthetic speech quality increases, especially for female speakers, when compared to integer delay predictors with the same number of taps.

The higher complexity of this type of predictor relative to the classical one is its major disadvantage. Suboptimal schemes in which the search for the fractional pitch delay is restricted to a neighbourhood of an integer pitch estimate can be envisaged, to decrease the computational load.

Although our tests were performed using an unquantized coder, designed for operation between 6 and 8 kbps, it is our belief that fractional pitch prediction will play a more dominant role at lower bit rates. The use of longer frame lengths, in this case, also suggests a further refinement of the predictor, in order to accommodate variable delay values in a single frame.

ACKNOWLEDGEMENT

We wish to acknowledge fruitful discussions with Peter Kroon, of AT&T Bell Laboratories.

REFERENCES

[1] J. Marques, J. Tribolet, I. Trancoso and L. Almeida, "Pitch Prediction with Fractional Delays in CELP Coding", Proc. EUROSPEECH, pp. 509-512, Paris, September 1989.

[2] I. Trancoso, L. Almeida and J. Tribolet, "A Study on the Relationships Between Stochastic and Harmonic Coding", Proc. ICASSP, pp. 1709-1712, Tokyo, May 1986.

[3] P. Kroon and E. Deprettere, "A Class of Analysis-by-Synthesis Predictive Coders for High Quality Coding at Rates Between 4.8 and 16 Kbit/s", IEEE J. Selected Areas in Comm., Vol. 6, No. 2, pp. 353-363, February 1988.

[4] B. Atal, "Predictive Coding of Speech at Low Bit Rates", IEEE Trans. Comm., Vol. COM-30, No. 4, pp. 600-614, April 1982.

[5] B. Atal and M. Schroeder, "Stochastic Coding at Very Low Bit Rates", Proc. Int. Conf. Commun., pp. 1610-1613, 1984.

[6] W. Kleijn, D. Krasinski and R. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP", Proc. ICASSP, pp. 155-158, New York, 1988.

[7] I. Trancoso et al., "CELP: a Candidate for GSM Half-Rate Coding", Proc. ICASSP, Albuquerque, April 1990.