PITCH PREDICTION WITH FRACTIONAL DELAYS IN CELP CODING


INESC, R. Alves Redol nº 9 - 29, P-1000 Lisboa Portugal

ABSTRACT

Time-domain coders use long-term predictors to exploit the quasi-periodic structure of voiced speech. However, one major drawback with classical predictors is the restriction of the delay to integer values, reducing the performance especially when the pitch period is short. This paper presents an extension of the prediction technique to encompass non-integer delays enabling a more accurate representation of quasi-periodic segments and an increased flexibility in the design of long-term predictors. The use of long-term predictors with fractional delays in CELP coding produces an enhancement of the harmonic structure in high frequencies and an improvement of quality for female speakers.

INTRODUCTION

Time-domain coders (e.g. Adaptive Predictive Coder, Multipulse Coder, CELP Coder) use long-delay predictors to exploit the quasi-periodic structure of voiced speech. The performance of pitch predictors is therefore an important factor to achieve high quality at low bit rates.

CELP coders [1] are presently the best candidates for low bit rate coding in the 4-6 Kbps bit rate. However, the quality of the synthetic speech produced by these coders is not very high, distortions being more pronounced in female voices. In general, we can say that synthetic speech produced by CELP coders sounds somewhat noisy and that this factor is mainly related to the performance of the coder in voiced segments. Spectrograms show that the synthetic speech has less harmonic structure than the original speech signal [4]. This means that the quasi-periodic structure of voiced speech segments is not being well modeled by the pitch predictor.

One major limitation of classical predictors is the restriction of the delay to integer values, reducing the performance of the predictor especially when the pitch period is short. This paper presents an extension of the prediction technique to non-integer delays. This extension enables us not only to fill the gaps between integers with fractional values, but also to use non-uniform (e.g. logarithmic) quantization of the delay, short delay values being more accurately encoded than long ones. The use of fractional delays is compatible both with classical (open-loop) implementations of the pitch predictor and with adaptive codebook (closed-loop) schemes [2].

PREDICTION WITH NON-INTEGER DELAYS

Voiced speech is produced by the vibration of the vocal folds exciting a time-varying system consisting of the vocal tract and including the radiation effects. Since the movements of the articulators (jaw, tongue, etc) are much slower than the interval between two glottal pulses, the response of the time-varying system to each pulse is highly correlated with the previous pulse responses and can therefore be predicted to a great extent from the past information. This has been the basis for the use of long-term predictors in time-domain coders.

A very common solution for long-delay prediction is the use of a single tap predictor

\[
sp(n) = G s(n-N)
\]  

(1)

where \(s(n)\) is the input signal, \(sp(n)\) is the predicted signal, \(N\) is the integer delay value and \(G\) is a gain computed by least squares. Since delays are restricted to integer values, the predictor is not able to cope with arbitrary pitch intervals without replacing in some way, the optimal delay by an integer value, adding a "jitter" component and degrading the performance of the predictor in terms of SNR.

To overcome this difficulty we have tried to extend the concept of single tap predictor to encompass non-integer (fractional) delays. However, expression (1) is no longer valid if the integer delay \(N\) is replaced by an arbitrary delay \(D\), since the discrete signal \(s(.)\) is only defined for integer values of the argument. To allow the use of non-integer delays, the discrete signal has to be interpolated between sampling instants, i.e., we can still maintain an expression similar to (1)

\[
sp(n) = G s(n-D)
\]  

(2)

provided that prediction is based on the continuous-time signal \(s_c(t)\) which coincides with the discrete signal at the sampling instants.

Assuming that the continuous-time speech signal is sampled according to the Nyquist criterion, it can be recovered by the expression

\[
s_c(t) = \sum_{k=-\infty}^{\infty} s(k) \cdot \text{Sa}(\pi(t-k))
\]  

(3)
where $S_a(x)$ is the sampling function

$$S_a(x) = \frac{\sin x}{x}$$

(4)

Therefore, the predicted value is given by

$$s_p(n) = G \sum_{k=-\infty}^{\infty} S_a(\pi(n-D-k))$$

(5)

In practice, however, the sum is truncated to a finite number of terms and the predictor becomes

$$s_p(n) = G \sum_{k=P-M}^{P+M} S_a(\pi(n-D-k))$$

(6)

where $P$ denotes the nearest integer to $n-D$. Unfortunately, even this expression can not always be used because we sometimes do not know all the input values from $P-M$ to $P+M$. In frame by frame techniques, for instance, it is typical to only know the input until the beginning of the present frame $n_b$ (figure 1). To predict the $n$-th sample the best we can do is to use the input values from $-\infty$ to $n_b-1$ and the predicted values from $n_b$ to $n-1$. Therefore, the predictor expression takes the form

$$s_p(n) = G \sum_{k=P-M}^{P+M} S_a(\pi(n-D-k)) + G \sum_{k=\min(P+M,n_b)}^{\min(P+M,1+n_b)} s_p(k)$$

(7)

It is apparent from figure 1 that for sufficiently long-delays, expression (7) reduces to equation (6).

![Diagram of delay estimation](image)

**Figure 1 - Data used to predict the n-th sample with delay D**

### DELAY ESTIMATION

Let us now address the problem of delay estimation. Delay is usually evaluated by minimization of a quadratic cost functional

$$J(D) = \sum_{n=n_b}^{n_b+L-1} [s(n) - s_p(n;D)]^2$$

(8)

where $n_b$ is the beginning of the frame, $L$ is the length of the optimization interval and $s_p(n;D)$ is the output of the predictor with delay $D$. Since the gain $G$ of the predictor is also obtained by the minimization of cost functional $J(D)$ through the expression

$$G(D) = \frac{\sum_{n=n_b}^{n_b+L-1} s(n) s_p(n;D)^2}{\sum_{n=n_b}^{n_b+L-1} s_p(n;D)^2}$$

(9)

where $s_p(n;D)$ denotes the normalized predictor computed with unitary gain, the minimization of the quadratic cost functional is equivalent to the maximization of the functional

$$C(D) = \frac{\sum_{n=n_b}^{n_b+L-1} s(n)^2 s_p(n;D)^2}{\sum_{n=n_b}^{n_b+L-1} s_p(n;D)^2}$$

(10)

The choice of the best delay value consists therefore of the evaluation 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 choosing the largest value. The same procedure will be used in this paper to deal with non-uniform discretization of the delay. However, a criticism applies to both:

i) 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 sufficient to characterize the function.

ii) Even if $C(D)$ is sampled according to the Nyquist criterion the maximum of the sampled signal can be located near local maxima of the continuous-time signal instead of being in the neighbourhood of the global maximum.

Figure 2 clarifies this point by showing the LPC residue of a voiced speech segment and the function $C(D)$ used to evaluate the optimal delay for the long-delay prediction of the residual signal, for the frame $t = 20$ ms to $t = 27.5$ ms. The speech signal was sampled at 8 KHz and the function $C(D)$ shown in figure 2b was evaluated with 8 fractional delays per sampling interval. Figure 2c displays 8 down-sampled versions of figure 2b 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

$$D^j = \frac{1}{64} j \text{ ms} \quad j=0,1,2, ..., 7$$

(11)

It is concluded from these 8 down-sampled versions of $C(D)$ shown in figure 2c that depending on the delay grid used to evaluate the

\[\text{Figure 2: Data used to predict the n-th sample with delay D}\]

\[\text{Delay Estimation}\]

\[\text{Figure 2: Data used to predict the n-th sample with delay D}\]

\[\text{Delay Estimation}\]
function C(D) we will choose an optimal delay corresponding to one
pitch interval (1st function), two pitch intervals (3rd function) or
even 3 pitch intervals (5th function). This mechanism may therefore
be responsible for the occurrence of many delay jumps in long-delay
prediction of speech.

![LPC residue](image)

![C(D)](image)

Figure 2 - Delay Estimation: a) LPC residue; b) function C(D); c) down-sampled versions of C(D)

Although we have put some stress on the dangers of a least squares
estimation of the delay when the delay grid is sparse, we still did use
this technique to test the performance of long-delay prediction with
fractional delays and compare it with integer delay predictors.
However, although we maintain the same delay estimation
algorithm, the replacement of the integer delay grid by a non-uniform
grid with fractional delays can have a strong influence on the delay
estimates. Figure 3 shows an example where we compare the delay
evolution in a female speech utterance using pitch prediction with
integer and fractional delays. In figure 3a, the optimal delay is
evaluated using a grid of 128 integer values ranging from 2.5 to 18.5 ms.
We can distinguish the pitch contour and a large number of jumps to multiples of the pitch period.

![a](image)

![b](image)

Figure 3 - Evolution of the Optimal Delay in Long-Term Prediction
a) with integer delays; b) with fractional delays

In figure 3b the same delay range was filled with 256 fractional
values logarithmically distributed in the same interval, leading to a
much denser grid for short delays. The effect on the delay
trajectory, shown in figure 3b, was a significant reduction on the
number of jumps and therefore a much smoother trajectory. This is
partially explained by the fact that some jumps in figure 3a 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 addition, we
have used in figure 3b a grid of delay values which is denser for
short delays than for longer ones enabling a better detection of the
maxima of C(D) for short delay values.

APPLICATION TO CELP CODING

Let us now consider the use of long-term prediction with fractional
delays in the context of CELP coding. In previous sections the issue
of the appropriate sampling of the delay values for long-delay
prediction was raised. The approach is reminiscent of vector
quantization techniques since the delay is selected from a codebook
of possible candidates by the minimization of a distortion measure.
The introduction of long-term prediction with fractional delays has
allowed a great amount of flexibility in the design of the delay
codebook, which can ultimately be obtained by a training procedure.
Finally we showed the dangers of an inappropriate sampling which
can lead to many delay jumps especially when the quantization grid
is sparse for short delay values. These jumps and the jitter associated
with sparse delay quantization produce distortions on the short-time
spectrum which destroy the harmonic structure mainly at high
frequencies, being perceived as noise in listening tests. Traditionally,
2 or 3 tap long-term predictors have been used in time-domain
coders in order to provide a frequency-dependent gain factor and an
"interpolated" value for the non-integer delay and thus improve the
synthetic speech quality especially at high frequencies. However,
one can use fractional-delay predictors for this purpose without
having to quantize multiple gain coefficients.

Several directions can be followed to improve long-term prediction
using the concept of fractional delays. However, we will restrict
the discussion to the study of non-uniform delay quantization. Three
basic schemes will be compared in a CELP coding framework: long-
term prediction with integer delays (with 1 and 2-taps), and the
fractional delay predictor with logarithmically quantized delays. The
first two predictors use 128 integer delay values ranging from 2.5 to
18.5 ms while the third predictor uses 256 fractional delays
logarithmically distributed in the same range. Each of the three
prediction schemes was tested in open and closed-loop strategies.

The comparison was performed using an unquantized CELP
algorithm similar to the reference coder described in [3]. 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 algorithm was then applied to a
small data base consisting of 16 speech utterances (8 uttered by male
speakers and 8 female by female speakers) corresponding to a total
of 45 s of speech.

Figure 4 shows spectrograms of original and synthetic voiced speech utterances produced by the CELP algorithm. Figure 4b shows the results obtained with a single tap predictor with integer delays. One can easily notice the increase of noise between the harmonics in the lower region of the spectrum and the destruction of the harmonic structure in higher frequencies. Figure 4c displays the results obtained with fractional delays which enable a better representation of the harmonic structure above 2 KHz.

![Spectrograms](a) original; (b) CELP with integer delays; (c) CELP with fractional delays

Table I summarises the segmental SNR results obtained for each of the 6 long-term predictors (3 open-loop predictors and 3 closed-loop predictors). For each predictor, we discriminate the SNR results for male speakers, for female speakers and for both male and female speakers and, in each case, we show the performance of the coder in voiced frames, in unvoiced frames and in whole utterances. One can conclude that fractional delay prediction has always improved the SNR results obtained with 1-tap integer delay prediction, yielding approximately the same SNR as two tap predictors. The objective improvement is more pronounced for female speakers in voiced frames.

Informal listening tests were performed to assess the performance of the prediction schemes. In female utterances, there was a clear preference for the results produced with fractional delays, over the results of the 1-tap predictor with integer delays, both in open and in closed-loop predictors. When compared with the 2-tap predictor, the single tap predictor with fractional delays was judged to be equivalent in the open-loop case, but worse in closed-loop. In male-speaker utterances, the differences between the performance of the three predictors were considerably smaller than in female speaker ones.

<table>
<thead>
<tr>
<th></th>
<th>open loop</th>
<th>closed loop</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i-1tap</td>
<td>i-2tap</td>
</tr>
<tr>
<td>male speakers</td>
<td>10.4</td>
<td>10.6</td>
</tr>
<tr>
<td>female speakers</td>
<td>12.0</td>
<td>12.5</td>
</tr>
<tr>
<td>male and female speakers</td>
<td>11.2</td>
<td>11.6</td>
</tr>
<tr>
<td>9.0</td>
<td>8.9</td>
<td>8.9</td>
</tr>
<tr>
<td>10.4</td>
<td>10.6</td>
<td>10.3</td>
</tr>
<tr>
<td>11.0</td>
<td>11.2</td>
<td>11.1</td>
</tr>
<tr>
<td>12.0</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>13.4</td>
<td>14.0</td>
<td>14.0</td>
</tr>
<tr>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
</tr>
</tbody>
</table>

Table I - Segmental SNR Results in CELP Coding

Performance of three long-delay predictors in open and closed-loop: i-1tap: integer delays with 1-tap; i-2-tap: integer delays with 2-taps; f-1tap: fractional delays with 1-tap

**CONCLUSION**

We have presented a long-term predictor with non-integer delays which enables an improved accuracy in the representation of voiced speech. The introduction of this predictor in a CELP algorithm produced an enhancement of the harmonic structure above 2 KHz and an improvement of the quality of the stochastic speech for female speakers. This strengthens our belief that the "roughness" which is heard in synthetic speech produced by CELP coders is closely related to the destruction of harmonic structure. This structure must therefore be enforced by an improvement of the long-term delay prediction mechanism or by post-filtering techniques [5].

**ACKNOWLEDGEMENT**

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

**REFERENCES**