Proc. IEEE Workshop Speech Coding (Ste. Adèle, QC), pp. 55-56, Oct. 1993

Pitch Filtering in CELP Using a Time Scaling Approach

Gebrael Chahine[1] and Peter Kabal[1,2]

[1] Department of Electrical Engineering
McGill University
Montreal, Quebec H3A 2A7

[2] INRS-Télécommunications
Université du Québec
Verdun, Quebec H3E 1H6

Abstract

In this paper, a new approach to pitch filtering in the synthesis stage of a Code-Excited Linear Prediction (CELP) coder is described. With this method, time scaling/shifting of the original speech is performed on a block by block basis in order to provide a better match between the pitch pulses in the original and reconstructed speech. This technique allows for a reduction of the pitch lag resolution, with the attendant bit rate reduction, of the CELP coder while maintaining the same speech quality.

1. Introduction

The pitch prediction stage in a Code-Excited Linear Prediction (CELP) coder contributes in a major way to good speech quality, especially at rates between 4 and 10 kbits/s. At high rates, a sufficient number of bits is assigned to the excitation signal enabling the coder to compensate for harmonic structure that the pitch synthesis filter fails to model. At low bit rates, the quality of the synthesized speech is much more dependent on the performance of the pitch filter. Improvements can be obtained by using a multi-tap pitch filter, but at the expense of requiring a greater number of bits to code the coefficient values. Alternately, a single-tap pitch filter can be improved by increasing the time resolution of the allowed pitch lags. This fractional delay pitch filter is better able to model pitch tracks and results in a better perceived periodicity, but at the cost of requiring more bits to represent the lag values.

We consider rates below 5 kbits/s, a rate at which the pitch filter performance begins to degrade as it becomes harder to recreate the evolution of the pitch cycle waveform with the restricted number of bits allocated to the pitch filter.

Consider a single tap pitch filter. The pitch filter is described by the pitch lag and pitch coefficient value, (M, β). The pitch parameters (M, β) and the excitation codebook parameters (index i and gain G) are determined in two steps. In the first step, G is set to zero. Let ϵ_p denote the minimum mean square error between the reference original speech signal and the reconstructed speech signal, and $(M_{\text{opt}}, \beta_p)$ be the corresponding delay and pitch coefficient. The basic idea of time scaling technique is to further minimize ϵ_p by allowing a better match between a modified original signal and the synthesized pitch structure.

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

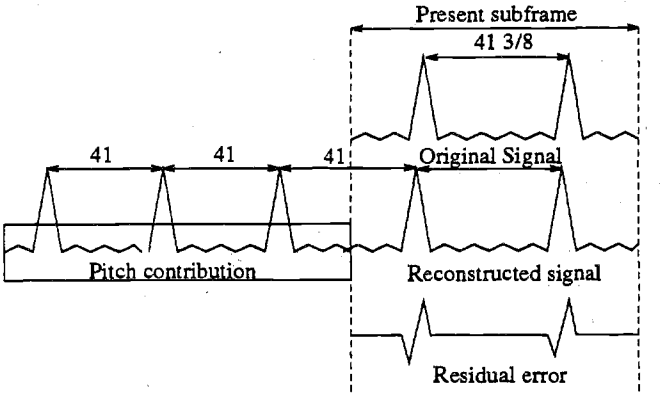

Fig. 1 Motivation for scaling/shifting.

2. Motivation for Time Scaling

Fig. 1 illustrates an idealized example. Let the reference original signal be periodic with a period of 41 3/8 samples. Assuming that only integer delay values are allowed, and that the pitch contribution is known, the reconstructed signal with a pitch lag of 41 will be as shown in Fig. 1. Note that the error shows vestiges of the pitch pulses. In the second stage search, the entry chosen from the excitation codebook tries to remove these pulses and thereby effectively shifting the pitch contribution to the correct position.

In the new scheme, the original signal is time scaled by the factor (41.375/41) to make the distance between the two pitch pulses equal in both the original and the reconstructed signals. A time shift is also needed in order to synchronize the pitch pulses. With this time scaling, the error between the signals decreases. In this way, the CELP codebook contribution no longer needs to compensate for mismatched pitch periods, but will instead concentrate on filling in the details not supplied by the pitch filter. With this scheme, the pitch pulses in the reconstructed speech are spaced differently and shifted with respect to the original pitch pulses. However, the aim is to make the scaling/shifting modifications of the original signal small enough that the perceived effect is negligible. As such, the scaling and shifting factors are not actually transmitted to the receiver.

In practice, the amount of time scaling and shifting must be determined by trial and error. A table containing combinations of time stretches/shrinks and shifts is used. This table is used to modify the original signal and if the modifications are small enough, there will be no effect on the perceived quality. Moreover, the size of the table is limited only by computational considerations. The modifications will be made independently block-by-block, whenever pitch parameters are updated. This allows for the easy retrofitting of this scheme into a conventional CELP coder.

3. Time Scaling Algorithm in CELP

In the first step in determining the pitch parameters, while M_{opt} is kept constant, an additional intermediate search is performed over all allowable time scaling/shifting values and pitch coefficients in order to determine the optimal pitch coefficient β_{opt} and the best time scaling/shifting parameters. Let the new minimum mean square error be denoted as ϵ_{min}.

The time scaling operation in the above algorithm corresponds to stretching or shrinking the original speech frames by a factor $w = M_{\text{opt}}/(M_{\text{opt}} + x)$, and an additional shift of s. The time scaling operation is done on a subframe basis ignoring continuity considerations. Large values of x or s can lead to a jittering of the pitch pulse locations. To counteract this, the allowable time scaling parameters x and s are kept small.

To accomplish the shifting and scaling, the problem consists of finding new sample values in between the original samples either by interpolation or extrapolation. A direct method is to digitally interpolate the signal to an extremely high frequency using a Finite Impulse Response (FIR) filter, and then choose the closest sample to the desired sampling instant. This method has been pursued by Ramstad [1] to include linear interpolation between the two interpolated samples which straddle the desired sample position. This scheme was used to scale and shift the signal.

4. Results and Conclusion

Using x and s limited to ± 1 and $\pm 5/100$ of a sample respectively, a direct implementation of the above algorithm does not perform as expected. We modified the scheme by noting that the larger shifts and scaling values are undesirable unless they result in a substantial decrease in the waveform match. A threshold, threshold, $t = 1 - |x|$, on the mean square error was introduced; the scaling operation is carried out only if $\epsilon_{\text{min}} < t\epsilon_p$.

Using the parameters for the FS-1016 CELP coder [2], the pitch delay M is coded with 8 bits capturing integer and fractional delays between 20 and 147 samples. Using our time scaling approach, we can achieve the same speech quality using only integer lags. This effectively saves one bit per subframe. This performance is achieved when $-0.4 \leq x \leq +0.4$ and s varying between $\pm 5\%$. Alternatively, by allowing only even-valued pitch delays (6 bits) to be used, using the new technique, a quality close to that for the integer delay case is obtained when $-0.8 \leq x \leq +0.8$. The objective measures (SNR, segmental SNR) are not good indicators of the performance of the time scaling/shifting approach because the original speech file is being modified during the synthesis parameters search. Subjective quality of the reconstructed speech is the preferred measure.

In conclusion, the time scaling approach allows for the reduction of the required pitch lag resolution while maintaining the quality of the reconstructed speech. No extra bits for side information are needed, as the time scaling information is not transmitted to the receiver.

References

[1] T. A. Ramstad, "Digital methods for conversion between arbitrary sampling frequencies", *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 32, pp. 577-591, June 1984.

[2] J. P. Campbell, T. E. Tremain and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder", *Digital Signal Processing, A Review Journal*, vol. 1, pp. 145–155, July 1991.