CASCADED LONG TERM PREDICTION FOR CODING POLYPHONIC AUDIO SIGNALS

Tejaswi Nanjundaswamy and Kenneth Rose

Department of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106-9560, USA
{tejaswi, rose}@ece.ucsb.edu

ABSTRACT

The long term prediction (LTP) tool is used in audio compression systems to exploit periodicity in signals. This tool capitalizes on the periodic component of the waveform by selecting a past segment as the basis for prediction of the current frame. However, most audio signals are polyphonic in nature, consisting of a mixture of periodic signals. This renders the LTP suboptimal, as the mixture's period equals the least common multiple of its individual component periods, which typically extends far beyond the duration over which the signal is stationary. Instead of seeking a past segment that represents a "compromise" for incompatible component periods, we propose a more complex filter that caters to the individual signal components. This proposed technique predicts every periodic component of the signal from its immediate history, and this is achieved by cascading LTP filters, each corresponding to individual periodic component. We also propose a recursive "divide and conquer" technique to estimate the parameters of all the LTP filters. For a real world evaluation, we employ this technique within the Bluetooth Sub-band Codec. Considerable gains achieved on a variety of polyphonic signals demonstrate the effectiveness of the proposal.

Index Terms— Audio coding, long term prediction, polyphonic signals

1. INTRODUCTION

The vast majority of speech and audio content consists of naturally occurring sounds which are periodic in nature. Examples include voiced parts of speech, music from string and wind instruments, etc. An audio signal with only one periodic component (i.e., a monophonic signal) obviously exhibits waveform repetition, which is exploited by the long term prediction (LTP) tool to improve compression efficiency. The tool essentially identifies a "similar" previous segment and scales it as the prediction for the current frame. The resulting low energy residue is encoded at reduced rate. The pitch period and scaling factor are either sent as side information or are backward adaptive, i.e., estimated from past reconstructed content at both encoder and decoder. The parameters are usually estimated via time domain waveform matching techniques that use a correlation measure to find a pitch period and scaling factor to minimize the mean squared prediction error. An LTP tool adaptation for the MPEG-4 Advanced Audio Coding [1] standard was proposed in [2].

While the existing LTP is well suited for content with a single periodic component, that is not the case for general audio which is often a mixture of multiple periodic signals. These signals belong to the class of polyphonic signals which includes as common examples, vocals with background music, orchestra, and chorus. Sound from a single instrument may also have multiple periodic components as in the case of the piano or the guitar. Of course, the mixture of periodic signals is itself periodic, but with a much longer period, namely, the least common multiple (LCM) of the individual component periods. The LTP tool would only be effective if the overall signal remained stationary over this longer period, which is rarely the case. Thus the LTP tool is suboptimal for polyphonic signals. It is nevertheless obvious that the redundancy implicit in the periodic components of the signal offers a significant potential for further gains, if exploited properly. An early investigation of approaches, based on similar underlying principles as we adopt here, was in the context of spectral estimation for sum of sinusoids [3]. Specifically, we propose cascading LTP filters corresponding to each periodic component in the mixture to form an overall *'cascaded long term prediction'* (CLTP) filter. This construct enables predicting each periodic component from the immediate past and only requires limited history.

For CLTP to be useful with real world signals it is obvious that an effective parameter estimation is of the utmost importance. To achieve this at acceptable complexity, while approaching optimality, we propose a "divide and conquer" recursive estimation technique. That is, we find optimal parameters of an individual filter in the cascade while fixing the parameters of the others. This process is then iterated for each filter in a loop, until convergence. Finally, we evaluate the proposed technique within a real-world coding system by integrating it into the ultra low delay Bluetooth Sub-band Codec (SBC) [4, 5]. As this system's capability to exploit redundancies is limited to a short frame length, an effective long term prediction has considerable potential impact on its performance. Experimental results substantiate the effectiveness of both the proposed CLTP paradigm and the optimization technique for a wide range of polyphonic signals.

2. POLYPHONIC SIGNALS AND PROBLEM SETTING

A simple periodic signal with pitch period N can be characterized via the relation $x[n] = x[n-N]$. But naturally occurring periodic signals are not perfectly stationary and have non-integral pitch periods. Thus a better characterization is

$$x[n] = \alpha x[n-N] + \beta x[n-N+1] \quad (1)$$

where α and β capture amplitude changes and approximate the non-integral pitch period via linear interpolation. A mixture of such periodic signals along with noise characterizes polyphonic audio

This work was supported by the NSF under grant CCF-0917230.

signal

$$s[n] = \sum_{i=0}^{P-1} x_i[n] + w[n] \quad (2)$$

where P is the number of periodic components, $w[n]$ is the noise sequence, and $x_i[n]$ are periodic signals following the general formula (1), i.e., satisfying

$$x_i[n] = \alpha_i x_i[n - N_i] + \beta_i x_i[n - N_i + 1]. \quad (3)$$

The prediction problem at hand is to find a filter of the form $H(z) = 1 - \sum_{k>0} a_k z^{-k}$ such that the prediction error $E(z) = S(z)H(z)$ is of minimum energy. If the signal has a single periodic component ($P = 1$), then we have an obvious choice for the LTP filter:

$$H_0(z) = 1 - \alpha_0 z^{-N_0} - \beta_0 z^{-N_0+1} \quad (4)$$

whose prediction error $e[n]$ is dependent only on the noise or innovation $w[n]$. If $P > 1$, the standard "compromise" LTP solution is

$$H_{\text{poly}}(z) = 1 - a' z^{-N_{\text{poly}}} - b' z^{-N_{\text{poly}}+1}. \quad (5)$$

where N_{poly} is the lag that minimizes the mean squared prediction error, within the history available for prediction. N_{poly} can be the LCM of individual periods $N_0, \ldots, N_{P-1}$, which, as discussed earlier, is suboptimal for real polyphonic signals as they do not remain stationary over this longer period. If LCM is beyond the history available then N_{poly} will be clearly a compromise that attempts to find a match despite the incompatible periods.

3. CASCADED LONG TERM PREDICTION

For polyphonic signals ($P > 1$), filtering with (4) results in

$$\begin{aligned} e_0[n] &= x_0[n] - \alpha_0 x_0[n - N_0] - \beta_0 x_0[n - N_0 + 1] + \\ &\quad \sum_{i=1}^{P-1} (x_i[n] - \alpha_0 x_i[n - N_0] - \beta_0 x_i[n - N_0 + 1]) \\ &\quad + w[n] - \alpha_0 w[n - N_0] - \beta_0 w[n - N_0 + 1]. \end{aligned} \quad (6)$$

Clearly, $x_0[n]$ is cancelled out, yielding

$$e_0[n] = \sum_{i=1}^{P-1} x'_i[n] + w'[n] \quad (7)$$

where $x'_i[n] = x_i[n] - \alpha_0 x_i[n - N_0] - \beta_0 x_i[n - N_0 + 1]$ and $w'[n] = w[n] - \alpha_0 w[n - N_0] - \beta_0 w[n - N_0 + 1]$ are the correspondingly modified version of the remaining periodic components, and noise, respectively. Straightforward algebra demonstrates the periodicity relationship of $x'_i[n]$:

$$\begin{aligned} x'_i[n] &= x_i[n] - \alpha_0 x_i[n - N_0] - \beta_0 x_i[n - N_0 + 1] \\ &= \alpha_i x_i[n - N_i] + \beta_i x_i[n - N_i + 1] - \\ &\quad \alpha_0(\alpha_i x_i[n - N_0 - N_i] + \\ &\quad\quad \beta_i x_i[n - N_0 - N_i + 1]) - \\ &\quad \beta_0(\alpha_i x_i[n - N_0 + 1 - N_i] + \\ &\quad\quad \beta_i x_i[n - N_0 + 1 - N_i + 1]) \\ &= \alpha_i(x_i[n - N_i] - \alpha_0 x_i[n - N_0 - N_i] - \\ &\quad\quad \beta_0 x_i[n - N_0 + 1 - N_i]) + \\ &\quad \beta_i(x_i[n - N_i + 1] - \alpha_0 x_i[n - N_0 - N_i + 1] - \\ &\quad\quad \beta_0 x_i[n - N_0 + 1 - N_i + 1]) \\ &= \alpha_i x'_i[n - N_i] + \beta_i x'_i[n - N_i + 1]. \end{aligned} \quad (8)$$

Hence $x'_i[n]$ exhibits the same periodicity as $x_i[n]$. Thus cascaded LTP filters

$$H_c(z) = \prod_{i=0}^{P-1} (1 - \alpha_i z^{-N_i} - \beta_i z^{-N_i+1}) \quad (9)$$

result in cancellation of all the periodic components and leaves a minimum energy prediction error, dependent only on $w[n]$. The CLTP filter of (9) forms the basis of our proposal to improve compression efficiency in polyphonic audio signals.

We note that the above analysis also suggests the potential usefulness of CLTP for separating periodic components. A survey of literature in this area, revealed a similar construct in early source separation work for multi-pitch estimation [6].

4. CLTP PARAMETER ESTIMATION

Estimation of CLTP filter parameter values is crucial for the effectiveness of this technique for real polyphonic signals. The complete parameter set includes P and N_i, α_i, β_i for $i = 0, \ldots, P-1$. A straightforward exhaustive approach for parameter value estimation would be to evaluate all combinations to find the one that minimizes the mean squared prediction error. This can be done by first fixing the number of periodic components P, then fixing the range of periods to Q possibilities, then finding the best α_i, β_i for each of the Q^P period combination and finally selecting the period combination which results in minimum mean squared prediction error. On repeating this for multiple P, we can select the best P. Clearly, the complexity of this approach grows exponentially with number of periodic components. Even on selecting a reasonable $Q = 100$ and $P = 5$, there are $\mathcal{O}(10^{10})$ combinations to be evaluated every time the parameters need an update, resulting in prohibitive computational complexity. Thus we propose a "divide and conquer" recursive estimation technique described below.

For a given P, to estimate parameters of the jth filter N_j, α_j, β_j we fix all the other filters and define

$$H_j(z) = \prod_{\forall i, i \neq j} (1 - \alpha_i z^{-N_i} - \beta_i z^{-N_i+1}) \quad (10)$$

$$S_j(z) = S(z)H_j(z). \quad (11)$$

We now find optimal parameters of the filter $\frac{H_c(z)}{H_j(z)} = 1 - \alpha_j z^{-N_j} - \beta_j z^{-N_j+1}$ in the intermediate output $s_j[n]$. This boils down to the classic LTP problem, where for a given N the α_j^N, β_j^N are given by

$$\begin{bmatrix} \alpha_j^N \\ \beta_j^N \end{bmatrix} = \begin{bmatrix} r_{(N,N)} & r_{(N-1,N)} \\ r_{(N-1,N)} & r_{(N-1,N-1)} \end{bmatrix}^{-1} \begin{bmatrix} r_{(0,N)} \\ r_{(0,N-1)} \end{bmatrix} \quad (12)$$

where the correlation values $r_{(k,l)}$ are

$$r_{(k,l)} = \sum s_j[n - k] s_j[n - l]. \quad (13)$$

And the optimal N_j is found as

$$N_j = \underset{N \in [N_{\min}, N_{\max}]}{\arg\min} \sum \begin{pmatrix} s_j[n] - \alpha_j^N s_j[n - N] \\ - \beta_j^N s_j[n - N + 1] \end{pmatrix}^2 \quad (14)$$

where $N_{\min}, N_{\max}$ are the lower and upper boundaries of the period search range. The range of summations in equations (13) and

(14) depend on whether the parameters are backward adaptive or not. This process is now iterated over the component filters of the cascade, until convergence. Convergence is guaranteed as the overall prediction error is monotone non-increasing at every step of the iteration. Note that as the overall cost has a clear non-convex dependency on pitch periods N_j, a globally optimal solution cannot be guaranteed by this technique. To optimize P, this process is repeated for consecutive values of P, starting at $P = 0$ (i.e., no prediction) to the upper limit of $P_{\max}$, and the combination that minimizes the prediction error energy is the complete set of CLTP parameters.

5. INTEGRATION WITH A REAL WORLD CODEC: BLUETOOTH SBC

Bluetooth Sub-band Codec (SBC) [4, 5] is a very simple ultra-low-delay compression technique proposed for short range wireless audio transmission. The SBC encoder first analyzes the audio signal into $B \in \{4 \text{ or } 8\}$ sub-bands and then blocks of $K \in \{4, 8, 12 \text{ or } 16\}$ samples in each sub-band are quantized adaptively. The quantization step sizes and the quantized data forms the bitstream received at the decoder, which dequantizes the sub-band content and then synthesizes the output signal from the sub-bands. We can clearly see that SBC's capability to exploit redundancies is limited only to small block lengths. Thus CLTP can improve its compression efficiency by providing effective inter-frame prediction, *without increasing delay*. The proposed CLTP parameter estimation technique is also well matched with the quantizer in SBC as both minimize MSE. We specifically employ CLTP in the first sub-band of SBC, before the quantization module, to exploit inter-block correlation. Operating only in the first sub-band minimizes tool complexity, while offering the bulk of the gains via effective prediction of the critical low frequencies.

Let $M = 512$ be a parameter that specifies the "look back" or amount of past reconstruction used for prediction parameter estimation. The first sub-band samples form $s[n], n = 0, \ldots, K-1$, and are predicted from previously reconstructed samples, $\hat{s}[n], n = -4M/B, \ldots, -1$, which are available at both the encoder and decoder. Given P, the parameters N_j, α_j, β_j are estimated, once per frame, backward adaptively via the recursive technique described in the previous section, where the summation in (13) and (14) consist of M/B terms and $N_{\min} = 100/B$, $N_{\max} = 800/B$. The parameter $P \in \{0, \ldots, P_{\max}\}$ is selected to minimize the mean squared prediction error. To estimate this error, the previously reconstructed samples $\hat{s}[n]$ are used as initial state for a given P's synthesis filter $1/H_c(z)$ to generate predicted samples $\tilde{s}[n], n = 0, \ldots, K-1$. Finally, the selected P is sent as side information to the decoder, with the quantized prediction error of first sub-band, the quantized samples of other sub-bands and the quantization step sizes.

The decoder receives P and estimates N_j, α_j, β_j to generate the predicted first sub-band samples. Then the quantized residue is added to generate the reconstructed first sub-band samples. The other sub-bands are reconstructed as in the current SBC decoder and finally, the output signal is synthesized from the sub-bands. To improve the recursive technique's rate of convergence, we use prediction parameters of the previous frame as initialization for the current frame.

We note that while the use of backward adaptive parameters reduces the side information rate, it also adds significant complexity to the decoder. As Bluetooth decoders are usually low power devices, we are pursuing, as a future direction, the use of differentially encoded forward adaptive parameters, so as to eliminate the need for parameter estimation at the decoder, at the cost of modest increase in side information.

	Prediction gains		Reconstruction gains	
Filename	LTP	CLTP	LTP	CLTP
Piano	5.8	15.0 (+9.2)	3.2	6.9 (+3.7)
Guitar	9.5	15.9 (+6.4)	5.0	7.9 (+2.9)
Harp	6.5	14.4 (+7.9)	5.8	12.6 (+6.8)
Bells	6.0	16.7 (+10.7)	5.4	13.9 (+8.5)
Mfv	11.6	19.0 (+7.4)	11.5	16.8 (+5.3)
Mozart	7.9	15.4 (+7.5)	6.3	11.5 (+5.2)
Quartet	3.0	7.3 (+4.3)	2.3	5.7 (+3.4)
Average	7.2	14.8 (+7.6)	5.6	10.8 (+5.2)

Table 1: Prediction gains and reconstruction gains in dB

6. RESULTS

In our experiments, we compare the following coders:

- Reference SBC with no prediction (referred to in figure as "NoLTP")
- SBC with a single LTP filter (obtained by setting $P_{\max} = 1$)
- SBC with the proposed CLTP.

The SBC is operated at $B = 4$ and $K = 16$ and for CLTP $P_{\max} = 5$. Thus side information rate is 1 bit/block for LTP (0.7/0.75 kbps) and 3 bits/block for CLTP (2.1/2.25 kbps) and are included in the rate totals. The experiments are conducted with single channel 44.1/48kHz audio sample subset from the standard MPEG and EBU SQAM database. We select only a 4 seconds portion of each audio file to reduce computation and evaluation times. The resulting subset is:

- Single instrument multiple chords: Grand Piano, Guitar, Harp, Tubular Bells
- Orchestra: Mfv, Mozart
- Chorus: Vocal Quartet

As SBC encodes with the aim of minimizing signal to quantization noise ratio (MSE criteria), we evaluate SNR gains to measure our performance improvements. The prediction gains and the reconstruction gains at an operating point of around 100 kbps, for each of the seven files, are given in Table 1. The table shows that CLTP provides truly major prediction gains of on the average 7.6 dB over LTP, which translate to substantial compression performance gains of on the average 5.2 dB. The table also shows that despite the significant gains already provided by LTP, it nevertheless left much room for improvement. The variability in how much of the CLTP prediction gains are retained may be attributed to the fact that prediction is done only in the first sub-band, whose relative importance varies across signal types.

We then evaluate SNR versus bitrate to generate operational rate-distortion (RD) plots for each coder. RD plots averaged over the test dataset of seven files, are shown in Figure 1. The plots

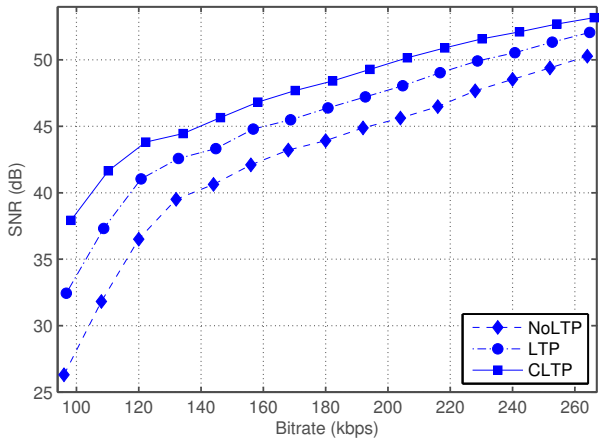

Figure 1: Signal to quantization noise ratio versus bit rate of the competing coders, evaluated and averaged over the seven files

clearly demonstrate the substantial gains provided by CLTP for a wide range of polyphonic signals. Informal listening tests confirm that the significant gains in objective criteria translate to substantial subjective quality improvements as well. More comprehensive subjective testing will be provided in future work that extends the applicability of CLTP to perceptual coders such as MPEG-AAC.

Note that the proposed technique is of higher complexity than LTP. The increase is mainly attributed to the parameter estimation done at each P, recursively. As the main objective of this work was to validate the concept of CLTP, no significant effort was put into minimizing complexity. Without complexity optimization, a crude implementation of the proposed encoder took on the average 130 times longer than LTP for the evaluated dataset, wherein the recursive technique took on the average 25 iterations to converge. It is clear that there are many simple ways to drastically reduce the complexity (all beyond the scope of this paper) including, for example, controlling the convergence criteria to optimize the tradeoff between complexity and prediction quality. Also note that the encoder with conventional LTP was 15 times more complex than encoding without prediction.

7. CONCLUSION

This paper developed a novel cascaded long term prediction filtering approach for improving the compression efficiency for polyphonic audio signals. Compared to the existing LTP technique, which is limited to predicting the mixture of periodic signals via a compromise in terms of a shared lag, the proposed technique predicts individual components optimally from their immediate past. We also proposed an effective, recursive technique for estimation of the filter parameters. The approach was applied within the real world compression system of the Bluetooth Sub-band Codec, and the results show considerable gains that substantiate the effectiveness of the approach in exploiting redundancies within mixtures of periodic signals. Future directions include extending CLTP to perceptual coders such as MPEG AAC.

8. REFERENCES

[1] ISO/IEC 14496-3:2005, "Information technology - Coding of audio-visual objects - Part 3: Audio - Subpart 4: General audio coding (GA)," 2005.

[2] J. Ojanperä, M. Väänänen, and L. Yin, "Long term predictor for transform domain perceptual audio coding," in *Proc. 107th AES Convention*, Sep. 1999, paper 5036.

[3] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[4] Bluetooth Audio Video Working Group, *Bluetooth Specification: Advanced Audio Distribution Profile*. Bluetooth SIG Inc., 2002.

[5] F. de Bont, M. Groenewegen, and W. Oomen, "A high quality audio-coding system at 128 kb/s," in *Proc. 98th AES Convention*, Feb. 1995, paper 3937.

[6] A. de Cheveigné, "A mixed speech F_0 estimation algorithm," in *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech '91)*, Sept. 1991.