BIDIRECTIONAL CASCADED LONG TERM PREDICTION FOR FRAME LOSS CONCEALMENT IN 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

This paper proposes a frame loss concealment technique for audio signals, which is designed to overcome the main challenge due to the polyphonic nature of most music signals and is inspired by our recent research on compression of such signals. The underlying idea is to employ a cascade of long term prediction filters (tailored to the periodic components) to circumvent the pitfalls of naive waveform repetition, and to enable effective time-domain prediction of every periodic component from the immediate history. In the first phase, a cascaded filter is designed from available past samples and is used to predict across the lost frame(s). Available future reconstructed samples allow refinement of the filter parameters to minimize the squared prediction error across such samples. In the second phase a prediction is similarly performed in reverse from future samples. Finally the lost frame is interpolated as a weighted average of forward and backward predicted samples. Objective and subjective evaluation results for the proposed approach, in comparison with existing techniques, all incorporated within an MPEG AAC low delay decoder, provide strong evidence for considerable gains across a variety of polyphonic signals.

Index Terms— frame loss concealment, long term prediction, polyphonic signals

1. INTRODUCTION

Audio transmission over networks enables a wide range of applications such as multimedia streaming, online radio and high-definition teleconferencing. These applications are often plagued by the problem of unreliable networking conditions, which leads to intermittent loss of data. Frame loss concealment (FLC) forms a crucial tool amongst the various strategies used to mitigate this issue. The FLC objective is to exploit all available information to approximate the lost frame while maintaining smooth transition with neighboring frames.

Various techniques have been proposed for FLC, amongst which the simple techniques of replacing the lost frame with silence or the previous frame, result in poor quality [1]. Advanced techniques are usually based on source modeling and were inspired from solutions to the equivalent problem of click removal in audio restoration [2]. For example, speech signals have one periodic component, and FLC techniques based on pitch waveform repetition are widely used. But these techniques fail for most audio signals which are polyphonic in nature, because they contain a mixture of periodic components. In principle, the mixture is itself periodic with period equalling the least common multiple (LCM) of its individual periods, but the signal rarely remains stationary over this extended period, rendering the pitch repetition techniques ineffective. To handle signals with multiple periodic signals, various frequency domain techniques have been proposed. FLC techniques based on sub-band domain prediction [3, 4] handle multiple tonal components in each sub-band via a higher order linear predictor. This approach does not utilize samples from future frames and is effectively an extrapolation technique with the shortcoming that it disregards smooth transition into future frames. An alternative approach to perform FLC in the modified discrete cosine transform (MDCT) domain, which accounts for future frames, was developed in our group [5]. This technique isolated tonal components in MDCT domain and interpolated the relevant missing MDCT coefficients of the lost frame using available past and future frames. Its performance gains, while substantial, were limited in the presence of multiple periodic components in polyphonic signals, whenever isolating individual tonal components was compromised by the frequency resolution of MDCT. This problem is notably pronounced in low delay coders which use low resolution MDCT.

The shortcomings of existing FLC techniques motivated the approach proposed herein, which is inspired by our recent work on efficient compression of polyphonic signals [6], to predict each periodic component in the time domain from its immediate past. Specifically, a long term prediction filter corresponding to each periodic component is cascaded to form the *cascaded long term prediction* (CLTP) filter. A preliminary set of parameters for these filters is estimated from past reconstructed samples via a recursive divide and conquer technique. In this recursion, parameters of one filter in the cascade are estimated while parameters of the others are fixed, and the process is iterated until convergence. Amongst these preliminary parameters, the pitch periods of each component are assumed to be stationary during the lost frame, while the filter coefficients are enhanced via a multiplicative factor to minimize the squared prediction error across future reconstructed samples. The predicted samples required for this minimization are generated via the 'looped' prediction (described in [3]), wherein given all the parameters, the filter is operated in the synthesis mode in a loop, with predictor output acting as input to the filter as well. The minimization is achieved via the well known quasi-Newton method called limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method [7] along with backtracking line search [8] for step size. Similarly, another set of multiplicative factors are generated for predicting the lost frame in the reverse direction from future samples. Finally the two sets of predicted samples are overlap-added with a triangular window to reconstruct the lost frame. The proposed scheme is incorporated within an MPEG AAC low delay (LD) mode [9, 10] decoder, with band-wise

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

energy adjustment when there is a large deviation from the geometric mean of energies in the bands of adjacent frames. Subjective and objective evaluation results for a wide range of polyphonic signals substantiate the effectiveness of the proposed technique.

2. POLYPHONIC SIGNALS AND CLTP

The periodic components present in polyphonic signals cannot be characterized as $x[m] = x[m-N]$, as they usually do not have a integral period N or constant amplitude across periods. A realistic characterization is $x[m] = \alpha x[m-N] + \beta x[m-N+1]$, where α, β capture the non-integral period via linear interpolation and the change in amplitude. Many such periodic components put together with noise forms a polyphonic signal as,

$$x[m] = \sum_{i=0}^{P-1} x_i[m] + w[m], \tag{1}$$

where P is the number of periodic components, $w[m]$ is the noise sequence and each periodic component $x_i[m]$ satisfies $x_i[m] = \alpha_i x_i[m-N_i] + \beta_i x_i[m-N_i+1]$. The CLTP filter,

$$H_c(z) = \prod_{i=0}^{P-1} (1 - \alpha_i z^{-N_i} - \beta_i z^{-N_i+1}) \tag{2}$$

proposed in our recent publication for encoding polyphonic signals [6], clearly cancels out all the periodic components, by design. Such a CLTP filter also plays a central role in the FLC technique proposed here, but the filter is suitably modified to utilize all the information available for reconstructing a lost frame.

3. CLTP FOR FRAME LOSS CONCEALMENT

When a frame is lost and the CLTP filter is known, the samples of the lost frame are predicted by first padding the previously reconstructed samples by zeros and then operating the synthesis filter $1/H_c(z)$ in this region of zeros, while using the previously reconstructed samples as initial state. This type of technique was called 'looped' prediction in [3], wherein output samples are recursively fed back to the filter to generate future predicted samples. Clearly estimation of parameters is critical to the performance of this predictor and the FLC technique. The proposed parameter estimation method and details of the overall technique are described in the following subsections.

3.1. Estimation of preliminary set of CLTP parameters

Direct minimization of a squared error cost function between actual and predicted future samples, to estimate all the parameters of the CLTP filter, would be complex as this involves the step of 'looped' prediction. Thus we assume the signal to be quasi-stationary in the vicinity of the lost frame and estimate using the past reconstructed samples, the pitch period and a preliminary set of filter coefficients. This is achieved at acceptable complexity via a recursive "divide and conquer" technique as introduced in [6], which estimates parameters of one filter, with all the other filters fixed.

Let $\hat{x}[m], \; -M_p \leq m < 0$, be the M_p past reconstructed samples modeled as in (1) and available to the FLC module. For a given P, to estimate jth filter parameters N_j, α_j, β_j, the other filters are fixed and defined as,

$$\begin{aligned} H_j(z) &= \prod_{\forall i, i \neq j} (1 - \alpha_i z^{-N_i} - \beta_i z^{-N_i+1}) \\ \hat{X}_j(z) &= \hat{X}(z) H_j(z). \end{aligned}$$

To find the optimal parameters for the jth filter $\frac{H_c(z)}{H_j(z)} = 1 - \alpha_j z^{-N_j} - \beta_j z^{-N_j+1}$ we focus on the intermediate output $\hat{x}_j[m]$. This becomes the well known LTP problem and for a given N the $\alpha_{(j,N)}, \beta_{(j,N)}$ are given as,

$$\begin{bmatrix} \alpha_{(j,N)} \\ \beta_{(j,N)} \end{bmatrix} = \begin{bmatrix} r_{(N,N)} & r_{(N-1,N)} \\ r_{(N-1,N)} & r_{(N-1,N-1)} \end{bmatrix}^{-1} \begin{bmatrix} r_{(0,N)} \\ r_{(0,N-1)} \end{bmatrix}, \tag{3}$$

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

$$r_{(k,l)} = \sum \hat{x}_j[m-k] \hat{x}_j[m-l]. \tag{4}$$

Stability of the synthesis filter used for prediction is ensured by restricting $\alpha_{(j,N)}, \beta_{(j,N)}$ to satisfy the sufficient stability criterion:

$$|\alpha_{(j,N)}| + |\beta_{(j,N)}| \leq 1. \tag{5}$$

Clearly if $\alpha_{(j,N)}, \beta_{(j,N)}$ generated by (3) do not satisfy (5), then the best solution lies on boundary of the region, defined by the rhombus $|\alpha_{(j,N)}| + |\beta_{(j,N)}| = 1$. The new solution is obtained by finding parameters to the four filter combinations of $1 - \alpha_{(j,N)} z^{-N} \pm (1 \pm \alpha_{(j,N)}) z^{-N+1}$, then restricting values of $\alpha_{(j,N)}$ to be on the rhombus and selecting amongst these the one which minimizes the mean squared prediction error:

$$\varepsilon_N = \sum (\hat{x}_j[m] - \alpha_{(j,N)} \hat{x}_j[m-N] - \beta_{(j,N)} \hat{x}_j[m-N+1])^2. \tag{6}$$

Having determined $\alpha_{(j,N)}, \beta_{(j,N)}$ for every N, the remaining step is to find the optimal N_j via

$$N_j = \underset{N \in [N_{\min}, N_{\max}]}{\arg\min} \; \varepsilon_N, \tag{7}$$

where $N_{\min}, N_{\max}$ are the lower and upper limits of the period search range. The summation limits in (6) and (4) are determined by the operating frame length of the coder. The above process is iterated for each periodic component (and corresponding filter) until convergence. Convergence is guaranteed as each step of the iteration ensures monotonic decrease in the overall cost function.

3.2. CLTP parameter refinement

In the networking applications where FLC is mainly used, availability of future frames while reconstructing a lost frame is usually assured. That is, if a frame with K samples is lost, usually M_f future reconstructed samples given as $\hat{x}[m], \; K \leq m < K + M_f$, are available to the FLC module. Using these samples to reconstruct a lost frame that transitions smoothly into the future is critical for good concealment quality and this is achieved by refining the preliminary CLTP filter parameters. We nevertheless assume that the pitch periods N_i are stationary in the vicinity of the lost frame, and hence employ multiplicative factors G_i to form an updated CLTP filter,

$$H_c(z) = \prod_{i=0}^{P-1} (1 - G_i(\alpha_i z^{-N_i} + \beta_i z^{-N_i+1})). \tag{8}$$

The CLTP filter allows us to generate the predicted future samples $\tilde{x}[m], \; K \leq m < K + M_f$,via 'looped' prediction. We now adjust the multiplicative factors G_i such that they minimize the squared prediction error, i.e., the cost function is given as

$$\varepsilon(\mathbf{G}) = \sum_{m=K}^{K+M_f-1} (\hat{x}[m] - \tilde{x}[m])^2, \tag{9}$$

where $\mathbf{G} = [G_0, \ldots, G_{P-1}]$ is the set of all multiplicative factors. Since the cost function has a complex dependency on $\mathbf{G}$, we use a generic quasi-Newton optimization method called limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [7] method. This is chosen as it converges faster than a plain gradient descent method. More details about this iterative method can be found in [7]. Since calculating the gradient function of the cost function is also complex, we approximate the partial derivatives as a difference in cost function for a small perturbation, i.e.,

$$\frac{\partial \varepsilon(\mathbf{G})}{\partial G_i} \approx \frac{\varepsilon(\bar{\mathbf{G}}_\mathbf{i}, G_i + h) - \varepsilon(\bar{\mathbf{G}}_\mathbf{i}, G_i)}{h}, \tag{10}$$

where $\bar{\mathbf{G}}_\mathbf{i}$ is the set of all multiplicative factors except G_i. Also the step size used within the L-BFGS algorithm is adapted via the backtracking line search method described in [8]. We note that the cost function is not convex and thus the above optimization cannot guarantee a global optima. But, as we will see experimentally, locally optimal multiplicative factors provide substantial improvement in concealment quality as they adapt the prediction filter parameters to exploit the available future reconstructed samples. Given the resulting CLTP filter, one set of samples of the lost frame is generated via the 'looped' prediction as $\tilde{x}[m],\ 0 \leq m < K$.

3.3. Bidirectional prediction

Further improvement in concealment quality is achieved by using samples predicted in the reverse direction from the future samples. To use an approach similar to the one described above for prediction in the forward direction, a reversed set of reconstructed samples available to the FLC module, is defined as $\hat{x}_r[m] = \hat{x}[K-1-m]$. This set in the range $-M_f \leq m < 0$ forms the new "past" reconstructed samples and the range $K \leq m < K + M_p$ forms the new "future" reconstructed samples. Since pitch periods are assumed to be stationary close to the lost frame, we start with the same preliminary CLTP filter estimated in section 3.1 for the reverse direction as well and estimate a new set of multiplicative factors G_i^r via the technique described in section 3.2, to form the reverse CLTP filter,

$$H_c^r(z) = \prod_{i=0}^{P-1} (1 - G_i^r(\alpha_i z^{-N_i} + \beta_i z^{-N_i+1})). \tag{11}$$

Given this reverse CLTP filter, another set of samples of the lost frame is generated via the 'looped' prediction as $\tilde{x}_r[m],\ 0 \leq m < K$. Finally the overall lost frame $\tilde{x}_o[m],\ 0 \leq m < K$ is generated as a weighted average of the two sets as,

$$\tilde{x}_o[m] = \tilde{x}[m]g[m] + \tilde{x}_r[K-1-m](1-g[m]), \tag{12}$$

where $g[m] = (1 - m/(K-1))$ are the weights which are proportional to each predicted sample's distance from the set of reconstructed samples used for their generation.

3.4. Integration within MPEG AAC-LD

MPEG AAC-LD coder segments data into 50% overlapped frames of length $K = 1024$. Thus one frame data loss results in inability to reconstruct K samples. We use $M_p = 2K$ past reconstructed samples and $M_f = K/2$ future reconstructed samples. Note that to have $M_f = K/2$ future samples, 2 future frames have to be available to the FLC module. This requirement is same as the energy interpolation method proposed in [5], while more than the sub-band domain prediction based method proposed in [3], as it requires only one future frame to adjust energy within each band of the lost frame. The sub-band domain prediction based technique has been observed to result in poor concealment quality when compared to the method proposed in [5] and we hypothesize this to be due to the fact that the prediction does not account for smooth transition into future samples. Thus we emphasize on having future samples available and for this we need at least 2 future frames, as with only one future frame no future samples can be reconstructed due to the overlapped frames and the aliasing introduced during MDCT. Given the data of frame n is lost and the neighboring reconstructed samples are available, the FLC module first estimates the preliminary set of CLTP parameters via the method described in section 3.1 with the parameters $P = 3$, $N_{min} = 50$, $N_{max} = 800$ and equations (4) and (6) having summation terms from $-K/2$ to -1. Then these parameters are refined to account for future reconstructed samples as described in section 3.2 and one set of samples of the lost frame is generated. Another set is generated via prediction in the reverse direction from future samples and the overall reconstruction of the lost frame is obtained via the method described in section 3.3. These K reconstructed samples are now transformed into MDCT domain, which enables utilizing the aliased samples from adjacent frames for final reconstruction and also enables maintaining a smooth transition in energies between adjacent frames. For energy adjustment the MDCT coefficients are divided into scale-factor bands as described in the standard [9] and for each band l the energy in all three frames $e_n[l]$, $e_{n-1}[l]$ and $e_{n+1}[l]$ is calculated. Now energy in the reconstructed frame is corrected by comparing it with the geometric mean $e_{gm}[l] = \sqrt{e_{n-1}[l]e_{n+1}[l]}$ and a gain factor $f[l]$, which is multiplied with all MDCT coefficients of the band l, is calculated as,

$$f[l] = \begin{cases} \sqrt{\frac{e_{gm}[l]}{e_n[l]}}, & \text{if } \frac{e_n[l]}{e_{gm}[l]} > T \text{ or } \frac{e_n[l]}{e_{gm}[l]} < 1/T, \\ 1, & \text{otherwise.} \end{cases} \tag{13}$$

That is, if the energy in a band deviates a lot from the geometric mean of energies in corresponding bands of adjacent frames, then it is corrected to the geometric mean. The threshold is chosen as $T = 5$. After multiplying the MDCT coefficients with their corresponding gain factors, final time domain samples are generated via the inverse MDCT process.

4. RESULTS

In our experiments, MPEG reference AAC-LD encoder is operated at 64 kbps to generate the bit-streams and the following four decoder modes are compared:

- Reference decoder with no frame loss
- Reference decoder with sub-band prediction based FLC module as proposed in [3, 4] (further referred as SBP-FLC)
- Reference decoder with MDCT domain energy interpolation FLC module as proposed in [5] (further referred as MDCT-FLC)
- Reference decoder with the proposed CLTP based FLC module (further referred as CLTP-FLC)

For decoders operating with FLC module the frames were randomly dropped at the rate of 10%, with same frames dropped in every decoder for a fair comparison. Also for simplicity, loss of consecutive frames was not allowed. The sub-band prediction based FLC module was operated at best quality by deciding to switch to shaped noise insertion only after checking prediction gain in all 32 sub-bands. The experiments are conducted with 44.1/48 kHz single channel audio

Filename	SBP-FLC	MDCT-FLC	CLTP-FLC
Piano	-3.16	-0.67	5.10
Guitar	-1.95	0.19	7.15
Harp	-3.59	-1.77	3.80
Bells	-2.08	0.06	4.26
Mfv	2.27	0.34	11.53
Mozart	-2.03	1.22	8.4
Average	-1.76	-0.11	6.71 (+6.82)

Table 1. SSNR in dB for various FLC techniques

sample subset from the EBU-SQAM and MPEG dataset. We restrict the length of each test file to 4 seconds to reduce evaluation times. The test subset includes single instrument multiple chord files (Grand Piano, Guitar, Harp, Tubular Bells), and orchestra files (Mfv, Mozart).

We first evaluate segmental signal to noise ratio (SNR) as an objective measure. Segmental SNR (SSNR) is the average of SNR in dB at each of the lost frame. For SSNR the signal energy is of the originally decoded MDCT coefficients and noise energy is of the difference between originally decoded MDCT coefficients and the MDCT coefficients generated by an FLC module. SSNR results for each FLC technique, evaluated for all the files is given in Table 1. The table clearly shows that the lost frame reconstructed via the proposed FLC technique is closest to the original frame, with an average segmental SNR improvement of on the average 6.82 dB over previously known best technique described in [5]. Note that the poor SSNR results of the competitive methods is mainly because their objective is not to absolutely match the waveform of the lost frame and have sections of MDCT coefficients adjusted with random signs. Thus subjective evaluations were conducted to identify the true perceptual gains via the MUSHRA listening tests. The test items were scored on a scale of 0 (bad) to 100 (excellent) and the tests were conducted with 16 listeners. The tests compared the outputs of 3 FLC techniques along with a output decoded with no frame loss. Randomly ordered 6 versions of each audio sample were presented to the listeners and these were a hidden reference (Ref), a 3.5 kHz low-pass filtered anchor (Anc), decoder output with no frame loss (NoLoss), decoder outputs with SBP-FLC, MDCT-FLC and CLTP-FLC module with 10% frame loss. Figure 1 shows the results of these tests, which include the average MUSHRA scores and the 95% confidence intervals, for the two types of files. These subjective evaluation results clearly demonstrates the greatly improved quality due to the proposed FLC technique for a variety of polyphonic signals. Note that the proposed CLTP-FLC technique is of higher complexity, with its crude implementation being 70 times more complex than the SBP-FLC technique. Clearly there are many simple ways of complexity reduction, but they are all beyond the scope of this paper.

5. CONCLUSION

This paper demonstrates a novel bidirectional cascaded long term prediction based frame loss concealment technique which substantially improves the reconstruction quality for polyphonic signals when used with low delay coders. Contrary to the currently used frequency domain techniques, the proposed technique operates in time domain, but addresses the problem of multiple periodic components by cascading their corresponding LTP filters. The prediction is done in both directions to better utilize available future samples and the filter parameters in each direction are optimized to account for samples on the other side of the lost frame. Subjective and objective evaluation of the proposed technique deployed within MPEG AAC-LD decoder substantiates the effectiveness of the proposed technique. Future directions include enhancing the proposed technique to not assume pitch period to be stationary in the neighborhood of the lost frame.

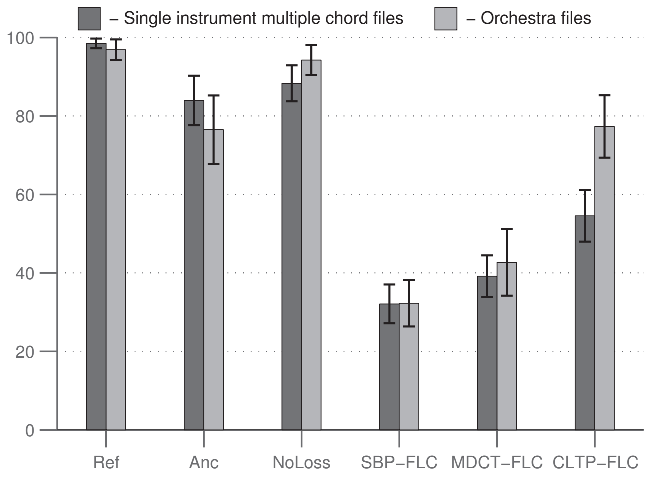

Fig. 1. MUSHRA listening test results comparing the FLC techniques

6. REFERENCES

[1] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[2] S.J. Godsill and P.J.W. Rayner, *Digital audio restoration: a statistical model based approach*, Springer verlag, 1998.

[3] J. Herre and E. Eberlein, "Evaluation of concealment techniques for compressed digital audio," in *Proc. 94th Conv. Aud. Eng. Soc*, Feb. 1993, Paper 3460.

[4] R. Sperschneider and P. Lauber, "Error concealment for compressed digital audio," in *Proc. 111th Conv. Aud. Eng. Soc*, Nov. 2003, Paper 5460.

[5] S.-U. Ryu and K. Rose, "An mdct domain frame-loss concealment technique for mpeg advanced audio coding," in *IEEE ICASSP*, 2007, pp. I–273–I–276.

[6] T. Nanjundaswamy and K. Rose, "Cascaded long term prediction for coding polyphonic audio signals," in *IEEE WASPAA*, Oct. 2011.

[7] J. Nocedal, "Updating quasi-newton matrices with limited storage," *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.

[8] J. Nocedal and S.J. Wright, *Numerical optimization*, Springer verlag, 1999.

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

[10] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "Mpeg-4 low delay audio coding based on the aac codec," in *Proc. 106th Conv. Aud. Eng. Soc*, May 1999, Paper 4929.