Time-Warping and Re-Phasing in Packet Loss Concealment

Robert Zopf, Jes Thyssen, and Juin-Hwey Chen

Broadcom Corporation, Irvine, California, USA
rzopf@broadcom.com, jthyssen@broadcom.com, rchen@broadcom.com

Abstract
This paper proposes two techniques to improve packet loss concealment (PLC). In the first technique, time-warping is used to stretch or shrink the time axis of the signal received in the first good frame after frame loss to align it with the extrapolated signal used to conceal the bad frame. This aligning procedure avoids any destructive interference that might otherwise occur when the two signals are out of phase and overlap-added. The second technique may be applied to speech codecs with memory, particularly suited for backward-adaptive systems. In this technique, called “re-phasing”, the internal states of the codec are phase-aligned with the signal in the first good frame. Both techniques are part of the ITU-T G.722 Appendix III packet loss concealment standard and provide significant quality improvement.

Index Terms: packet loss concealment, periodic waveform extrapolation, time-warping, re-phasing.

1. Introduction
Packet loss is a common problem in today’s ubiquitous packet networks. In order to conceal the degrading effects of the losses, a packet loss concealment (PLC) algorithm must be included at the decoder. For memoryless codecs such as ITU-T G.711, periodic waveform extrapolation (PWE) is a common approach [1] for PLC. For packet losses up to around 20 ms, this technique results in little or no audible distortion during the lost frames. The majority of the distortion occurs in the first good frame (FGF) where the PWE waveform must be merged with the received speech. To avoid discontinuity, the PWE waveform can be extended beyond the end of the bad frame and an overlap-add (OLA) operation with the FGF can then be performed. However, the true pitch period of the bad frame(s) in general does not follow the pitch track used during the waveform extrapolation. As a result, the extrapolated signal and the signal in the FGF may be out of phase and destructive interference can occur in the OLA region. If added delay is acceptable, the misalignment can be accounted for during the waveforms are in phase, whereas a positive value means that the two waveforms. A time lag of zero indicates that the two misaligned, an example of which is shown in Figure 1.

2. Time Lag Computation
The time-warping and re-phasing techniques require the number of samples that the PWE and FGF signals are misaligned, an example of which is shown in Figure 1.

Figure 1: Misalignment of FGF and PWE signals: (a) received signal, (b) PWE signal for lost frame and FGF.

The received signal with a 10 ms frame loss is shown in (a), while (b) is the PWE signal for the lost frame with extension into the FGF. In the FGF, the PWE signal leads the received signal by 1 ms. The time lag \( T_f \) is computed by searching for the peak of the normalized cross-correlation function \( R(k) \) between the signals for a time lag range of \( \pm \Delta_{\text{LAG}} \) around zero:

\[
R(k) = \frac{\sum_{i=0}^{L_w} s(i) \cdot \tilde{x}(i-k)}{\sqrt{\sum_{i=0}^{L_w} s^2(i) \cdot \sum_{i=0}^{L_w} \tilde{x}^2(i)}}, \quad k = -\Delta_{\text{LAG}}, \ldots, \Delta_{\text{LAG}}
\]

where \( s \) is the PWE signal, \( \tilde{x} \) is the FGF signal, \( L_w \) is the lag search window length, and \( i=0 \) represents the first sample in the lag search window. The time lag that maximizes this function corresponds to the relative time shift between the waveforms. A time lag of zero indicates that the two waveforms are in phase, whereas a positive value means that the FGF lags (is delayed compared to) the PWE and a negative value indicates the FGF leads the PWE.

The reliability of the time lag estimate is dependent upon the reconvergence rate of the signal in the FGF. This requires that the signal shape in the FGF recovers quickly enough to exhibit periodicity and provide a meaningful signal for the time lag calculation. For example, G.722 has this property.

of the mismatched internal states. In [3], the values for the forgetting factors are further modified and some improvement in quality is noted, but artifacts due to the state mismatch remain. To reduce the state mismatch problem, this paper introduces the concept of “re-phasing” which involves setting the internal states to a point in time where the PWE waveform is in-phase with the last input signal sample immediately before the FGF. Although re-phasing is mainly described in the context of backward-adaptive systems, it can also be used for forward-adaptive predictive coders.

The paper is organized as follows. Section 2 describes how the time lag between the PWE and FGF signals is computed. Time-warping and re-phasing are described in Section 3 and 4, respectively. Section 5 gives the details of how these techniques are used in ITU-T G.722 Appendix III [4] and presents some results. Finally, conclusions are given in Section 6.
3. Time-Warping

Time-warping is the process of stretching or shrinking a signal along the time axis. In order to maintain a continuous signal, the PWE signal and the FGF signal must be combined in a way that minimizes discontinuity. This is generally done by performing an OLA between the two signals. If the signals are out of phase with each other, waveform cancellation might occur and produce an audible artifact. This is illustrated in Figure 2 where the frame loss in (a) is applied to the received signal in (b) and PWE is applied in (c) with 2.5 ms of OLA at the start of the FGF (gray area). The pitch cycle waveform in (c) is different from that of (b) and causes audible distortion. The signal in (d) is a time-warped OLA that results in a waveform that more closely resembles the pitch cycle in (b).

As mentioned, the point at which the waveforms are in-phase should coincide with the center of the overlap-add window. The original lag estimate is computed over a relatively long window, the center of which may not coincide with the center of the OLA window. To improve alignment, a lag refinement search is performed. The search methodology is identical to that described above, except that (1) $L_d$ is set to the length of the OLA window, (2) the search is limited to a small range (such as $\pm 4$ samples), (3) the placement of the lag search window coincides with the expected OLA placement according to the original lag estimate, and (4) the extrapolated signal is offset by the original lag estimate.

There are many methods for performing the time-warping. The low-complexity approach used here involves a piece-wise single sample shift and overlap-add, starting from the end of the FGF. To perform shrinking, a sample is periodically dropped. From the point of sample drop, the original signal and the signal shifted right (due to the drop) are overlap-added. To perform stretching, a sample is periodically repeated. From the point of sample repeat, the original signal and the signal shifted to the left (due to the sample repeat) are overlap-added. The length of the overlap-add window depends on the periodicity of the sample add/drop. That is, adjacent windows should themselves not overlap. To limit complexity, the maximum window length is 8 samples.

The amount of time-warping was tuned to be constrained to $\pm 1.75$ ms for 10 ms frames. Warping by more than this may remove the destructive interference; but often introduces other audible distortion. In cases where the time lag is outside this range, no warping is done. Also, since the time-warping is performed within the FGF, frames beyond the FGF are unaffected. At a local level within the FGF, the search range of the time lag is symmetric; therefore, on average the time shift introduced into the FGF is zero.

4. Re-Phasing

The state mismatch between the encoder and decoder after frame loss for codecs with memory can cause significant artifacts in the decoded speech. This is especially true for backward-adaptive codecs. One method for updating the states after frame loss is to re-encode the PWE-based concealment waveform used in the bad frame [2]. Since the PWE signal is an estimate of the original signal, re-encoding it should update the states in a reasonable manner. For backward-adaptive codecs such as G.726 and G.722, it was observed that the state memory exhibits some degree of pitch modulation, and is hence, sensitive to the phase of the signal, especially if the frame boundary is near the pitch epoch where the signal magnitude rises and falls sharply. Because the phase lag used for generating the PWE signal may not exactly match the true pitch lag, the PWE and FGF signals may be out-of-phase at the frame boundary, thus the frame boundary may not be the most optimal time to stop re-encoding.

As an example, the state memory variable in G.722 exhibiting the greatest pitch modulation is the low-band log scaling factor, $NBL$:

$$NBL(n) = \beta \cdot NBL(n-1) + W_L[I_t(n-1)]$$

where $W_L$ is a logarithmic scaling factor multiplier, $\beta$ is a leakage constant, and $I_t$ is the truncated input codeword. The effect of phase misalignment on $NBL$ and the output signal is shown in Figure 4. The $NBL$ in error-free conditions and the corresponding segment of output speech is shown in (b). The pitch modulation of $NBL$ is clearly evident.

**Figure 2: Time-warped OLA:** (a) lost frame indicator, (b) error free, (c) simple OLA, (d) time-warped OLA.

Using time-warping, the original signal is time-warped in the direction towards the beginning of the frame to phase align the FGF signal with the PWE signal at some point in time within the first good frame. The amount of time-warping is controlled by the value of the lag $T_t$. It is desirable for the “in-phase point” to be in the middle of the OLA window, with the window positioned as close to the start of the frame as possible. This reduces the time in which the PWE signal is used as the PLC output signal. If the time lag is positive, the FGF signal is stretched and the OLA window can be positioned at the start of the FGF. However, if the lag is negative, the FGF signal is compressed and the OLA window will be positioned $|T_t|$ samples into the FGF.

Figure 3 is a zoomed-in version of the first good frame for the destructive interference case of Figure 2. Waveform (a) is the PWE signal, (b) is the FGF signal, and (c) is the FGF signal time-stretched from right to left by 1 ms to target alignment in the gray area. The OLA operation in the gray region between (a) and (b) is illustrated in (d) and exhibits cancellation, while the OLA operation between (a) and (c) shown in (e) maintains the pitch cycle waveform as seen in greater context in Figure 2(d).

**Figure 3: Alignment of FGF and PWE signals in the OLA region (grayed) by time-warping:** (a) PWE signal, (b) FGF signal without time-warping, (c) time-warped FGF signal, (d) OLA of (a) and (b), (e) OLA of (a) and (c).
output signal resulting from re-encoding the PWE signal is shown in (c) for the frame loss indicated in (a). The NBL greatly overshoots the true value immediately after the frame loss, resulting in a large artifact as visually evident in the output signal.

![Figure 4](image.png)

Figure 4: Effect of state memory misalignment and re-phasing on G.722 output: (a) lost frame indicator, (b) error free NBL (left) and output signal (right), (c) NBL (left) and output signal (right) with frame loss and no re-phasing, (d) NBL (left) and output signal (right) with frame loss and re-phased state memory.

To overcome this problem, the time lag is used to control where to stop the re-encoding process. Figure 5 illustrates the concept of re-phasing for the example in Figure 4.

![Figure 5](image.png)

Figure 5: Re-phasing of NBL

Under error-free conditions, the frame boundary occurs immediately preceding the sharp rise in NBL's magnitude. However, the PWE signal leads the original signal by 1.25 ms at the frame boundary; re-encoding PWE up to the frame boundary results in NBL taking the sharp rise in the lost frame update. Upon regular decoding in the FGF, NBL further increases, resulting in a large overshoot. The solution is to stop re-encoding PWE at a point that leaves the signals phase-aligned. Stopping the re-encoding at 8.75 ms allows the state memory to evolve more naturally as can be seen in Figure 4(d), where NBL's magnitude.

In the previous example, the PWE signal led the FGF signal, and re-encoding was halted before the frame boundary. If the PWE signal was to lag the FGF, the time-lag would be negative, and the number of samples to re-encode would be greater than the frame size (FS). In general, the number of samples (N) to re-encode is given by:

\[
N = FS - T_l
\]

The concept of re-phasing has been presented in the context of backward-adaptive predictive coding. However, it’s use is not limited to such. Most memory-based coders exhibit some phase dependency in the state memory and may benefit from re-phasing.

5. G.722 Appendix III Implementation

The time-warping and re-phasing techniques are part of ITU-T G.722 Appendix III. This PLC algorithm is a wideband speech domain PWE system with mixing of shaped noise based on the characteristics of the speech signal in frames preceding the frame loss [5]. The following subsections describe the time-warping and re-phasing in the PWE-based G.722 PLC, and then give a brief quality evaluation.

5.1. Low Complexity Re-Phasing and Time-Warping

To compute the time lag, the G.722 signal can be decoded in the FGF and used for correlation with the extrapolated signal. However, this brute force method is wasteful since the received stream will later be decoded using the re-phased signal. To avoid double decoding, the low complexity method illustrated in Figure 6 was developed.

![Figure 6](image.png)

Figure 6: Time-warping and re-phasing in G.722 Appendix III

During normal G.722 decoding, the low-band and high-band codes \( I_l(n) \) and \( I_h(n) \) are converted to difference signals, scaled by a backward-adaptive scale factor NBL and passed through a backward-adaptive pole-zero filter to obtain the sub-band signals which are then combined by a QMF synthesis filter bank to produce the output signal. At every sample in this process, the filter coefficients are updated. This update process accounts for a significant portion of the decoder complexity.

Since only a signal for lag computation is required, an approximation of the decoded FGF signal is sufficient. Thus, the coefficients of the two-pole, six-zero filter can be kept frozen when calculating the approximated FGF signal. In addition, since the lag is dependent upon the pitch, only a low-band approximation signal is derived. This low-band signal is fed to the Time-Lag Calculation block which computes the lag to a resolution of 8 kHz. Note that the low-band signal is not passed through the synthesis filter bank and does not contain the inherent 11.5-sample delay. This delay offset between the low-band signal and the full-band PWE signal needs to be accounted for in the alignment of the signals for the correlation calculation.

The initial state used for low-band approximation decoding is that obtained by re-encoding the PWE signal up to the frame boundary. Since this state is mismatched with
the encoder, the decoder output will take some time to recover. Though the signal magnitude and shape may still be converging, in practice the signal beyond 1 ms appears phase-aligned with the original signal and contains enough periodicity to reliably estimate the time lag. As such, only the signal beyond 1 ms is used in the time-lag calculation.

The time-lag $T_l$ is fed to the re-phasing module for re-encoding of the PWE signal and phase alignment of the state memory. Without re-phasing, the re-encoding could be done entirely in the bad frame. However, since the lag is not known until the GFG, the re-encoding cannot be completed in the bad frame. A simple approach would be to store the entire PWE signal and perform the re-encoding in the GFG. However, a more memory efficient and load balanced approach is used. The method performs re-encoding in the bad frame up to the frame boundary and preserves the internal state. In addition, the intermediate state after re-encoding FS-$\Delta_{\text{MAX}}$ samples is also stored. The PWE samples for re-encoding from FS-$\Delta_{\text{MAX}}$+1 to FS+$\Delta_{\text{MAX}}$ are saved in memory.

In the FGF, the low-band approximation decoding is performed using the stored state at the frame boundary as the initial state. If the lag is positive, the state at FS-$\Delta_{\text{MAX}}$ samples is restored and re-encoding commences for $\Delta_{\text{MAX}}$ + $T_l$ samples. If the lag is negative, the state at the frame boundary is used and an additional $T_l$ samples are re-encoded. At most, $\Delta_{\text{MAX}}$ samples are re-encoded in the GFG. G.722 decoding of the FGF is then performed using the re-phased state. Finally, this output is time-warped using a refined $T_l$ and then overlap-added with the PWE signal.

The techniques were implemented in fixed-point using the ITU-T STL2005 basic operators and consumed less than 1.5 WM0PS (excluding regular G.722 decoding).

### 5.2. Quality Evaluation

A total of 13 languages from the NTT-AT speech database each containing 96 sentence pairs of approximately 8 seconds were processed with and without re-phasing (time-warping was disabled in both cases). Random frame erasures were inserted at a rate of 5% with 10 ms frames. Wideband PESQ for both systems was calculated against the original speech.

The statistics of the sentence-wise PESQ difference between the systems with and without re-phasing is shown in Figure 7.

On the left side of Figure 7 the histogram of the 1248 PESQ differences is plotted. In order to improve the resolution of the outlier bins, the center bin is clipped. The histogram is positively skewed indicating the improvement when using re-phasing. On the right side, the average PESQ difference is plotted for PESQ differences with a magnitude greater than or equal to a threshold. In this way, small differences are excluded from the average and outlier performance can be evaluated. Note that the PESQ averages include both improvements and degradations in PESQ. It can be seen that as the threshold increases, the re-phasing is increasingly outperforming the system without re-phasing. About 7.3% of the sentence pairs have a PESQ difference greater than or equal to 0.1. For time-warping, it was found that PESQ penalizes the time-warped GFG (due to changing good samples in the FGF) and thus does not accurately reflect the perceptual improvement obtained by eliminating the waveform cancellation.

The improvements due to re-phasing and time-warping were also evaluated by expert listening. In the first evaluation, an expert listener evaluated a total of 13 one-minute files with 3% and 6% random and bursty pack loss. In total, time-warping and re-phasing were found to give audible improvement in 38 places, many of which were significant improvements, while they gave marginal degradation in only 3 places. In another experiment, a second listener evaluated 96 sentence pairs of 8 seconds each with 5% random packet loss and 10 ms frames. Three systems were evaluated and compared: no re-phasing or time-warping (System 1), re-phasing only (System 2), and both re-phasing and time-warping (System 3). The results are shown in Table 1.

### Table 1: Expert Listening Evaluation Results

<table>
<thead>
<tr>
<th>System Comparison</th>
<th>Improvements: Degradations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 versus 1</td>
<td>31:3</td>
</tr>
<tr>
<td>3 versus 2</td>
<td>28:4</td>
</tr>
<tr>
<td>3 versus 1</td>
<td>44:9</td>
</tr>
</tbody>
</table>

The first row indicates a clear improvement when re-phasing is included. The second row shows that significant further improvement is obtained when adding time-warping to a system with re-phasing. The third row shows a clear preference for applying both re-phasing and time-warping and is roughly consistent with the results of the first evaluation.

Re-phasing and time-warping were standardized as part of the PLC candidate selected by the ITU-T as G.722 Appendix III. This candidate was tested by independent subjective listening labs, and formal statistical analyses showed that it outperformed all other candidates.

### 6. Conclusion

This paper presented the time-warping and re-phasing techniques for improving packet loss concealment. These techniques address issues resulting from the misalignment of the waveform generated for concealment of the lost frame and the waveform decoded in the first good frame. A low complexity implementation, part of ITU-T G.722 Appendix III, was described as an example to show how these two techniques can be applied to the PLC of a codec. Various evaluations showed that these techniques provide a significant improvement in speech quality. In this paper, these techniques were mainly presented in the context of G.722 PLC; however, it should be emphasized that they are generally applicable to many other speech coders.

### 7. References