COMPRESSED DOMAIN PACKET LOSS CONCEALMENT OF SINUSOIDALLY CODED SPEECH

Christoffer A. Rødbro, Mads G. Christensen, Søren Vang Andersen, and Søren Holdt Jensen*

Department of Communication Technology,
Aalborg University, Denmark.
{car, mgc, sva, shj}@kom.auc.dk

ABSTRACT

In this paper we consider the problem of packet loss concealment for Voice over IP (VoIP). The speech signal is compressed at the transmitter using a sinusoidal coding scheme working at 8 kbit/s. At the receiver, packet loss concealment is carried out working directly on the quantized sinusoidal parameters, based on time-scaling of the packets surrounding the missing ones. Subjective listening tests show promising results indicating the potential of sinusoidal speech coding for VoIP.

1. INTRODUCTION

In packet-switched communication systems such as the Internet packets may be delayed or even lost during transmission. This is not critical in most applications since the receiving end can request retransmission of the packet in question. However, in a real-time constrained application such as VoIP, retransmission is not feasible since this would introduce a considerable delay prohibiting proper two-way conversation. Thus lost and delayed packets must be compensated for at the receiving end. This is usually attempted by storing a number of recently arrived packets in a jitter buffer before playout. If the packet delay is lower than the time extension of the jitter buffer it can be used to compensate for packet delay variations (jitter). However, packets delayed more than the length of the jitter buffer are considered lost and have to be replaced.

The simplest approaches in case of packet loss are silence or noise substitution but these methods have a high negative impact on perceived speech quality. Better approaches rely on waveform substitution from neighboring frames, see e.g. [1]. More recently, missing frames were estimated through a combination of LPC analysis and interpolation/extrapolation of the residual signal using sinusoidal modeling [2], [3]. Instead of estimating the missing packet, another approach is to stretch the packets preceding the missing one in order to allow more time for delayed packets to arrive [4], [5]. In a VoIP application the speech signal would normally be compressed to achieve a lower bit rate. An important design criterion for such speech coding schemes is robustness toward packet losses, see e.g. [6]. Moreover, the data made available by the speech coder at the receiver should be sufficient to facilitate packet loss concealment.

In this paper we utilize a speech coding algorithm based on sinusoidal modeling which is described in Sec. 2. In Sec. 3 we then propose a packet loss concealment algorithm based on time-scale modification which works directly on the sinusoidal parameters. The sinusoidal coding scheme is a modified version of that presented in [7] whereas the packet loss concealment is based on [8]. Experimental results are presented and discussed in Sec. 4 before Sec. 5 concludes on the work.

2. SINUSOIDAL CODER

Speech coding for use in packet switched networks should be designed for robustness toward packet losses. One way of achieving this is to ensure that decoding of frames can be performed independently. Also, it is desirable to design the coder in such a way that it is possible to perform packet loss concealment in the compressed domain. These properties can easily be incorporated into a sinusoidal coder.

We have developed a fixed bit-rate sinusoidal coder operating at 8 kbit/s suitable for packet switched networks as a reference system for testing the packet loss concealment method proposed. This is done to ensure that the method can operate under realistic conditions with quantized parameters.

The coder of [7] has been modified to fit the requirements of packet switched networks. It is based on a harmonic sinusoidal model, where the speech segment is represented as a finite sum of harmonically related sinusoids:

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l) \tag{1}$$

*Part of Mads G. Christensen's work was conducted within the ARDOR project, supported by the EU grant no. IST–2001–34095.

Here ω_0 is the fundamental frequency and L is the number of components in the segment, and A_l and ϕ_l are the amplitude and phase of the l'th harmonic respectively. After segmentation the parameters of this model are estimated. Here, the speech is split into segments of 20 ms with 50% overlap.

First, the pitch is estimated using the correlation based method proposed in [9]. The problem of finding the optimum amplitudes and phases then turns into a linear least-squares problem that is solved using weighted least squares (WLS), see e.g. [8] for details.

Although the harmonic sinusoidal model is only physiologically founded for voiced speech, it is well-known that it can be used for modeling of noise-like signals [10] such as unvoiced speech, provided that the frequency spacing is sufficiently small. A frequency spacing of 100 Hz for unvoiced speech has been found to form a reasonable tradeoff between model performance and the number of parameters. The cumulative mean normalized difference function in [9] is used for voiced/unvoiced decision and to estimate a voicing dependent cut-off frequency, ω_c.

The amplitudes are represented using a 10th order discrete all-pole model [11]. In this model the spectral envelope is optimized to match only at the discrete harmonic frequencies rather than the continuous spectrum. It is then represented using line spectral frequencies and finally "transparently" coded using perceptually weighted split vector quantization with a 24 bit codebook as described in [12].

The fundamental frequency and the gain are quantized in the log-domain using 7 and 5 bits respectively.

The phases can be represented efficiently by exploiting the near-linear relationship between the phases of the harmonics of voiced speech. This has been done by fitting a line to the unwrapped phases and the parameters of the line are encoded using a total of 7 bits. As the phases are only approximately linear and only in perfectly voiced regions, there are non-zero phase residuals or errors. These are then quantized using a scalar uniform quantizer in the range $]-\pi,\pi]$. Bits are allocated in accordance with the power distribution (the quantized DAP envelope) such that the phases of the largest components receive more bits than smaller ones. In unvoiced regions the phases are simply quantized directly. The reason for using bits for phase quantization in unvoiced segments is that it provides better modeling as waveform approximating capabilities are achieved. This is important in e.g. the burst of a plosive, where the phases are not stochastic. Also, it has been found to generally improve the perceived quality as well as improving robustness due to the waveform approximating property.

In Table 1 the bit allocation per frame of the coder for operation at 8 kbit/s is shown. In the decoding process phase randomization inversely proportional to the number of bits allocated for a given component should be applied with different ranges depending on the voicing of the components to avoid unnatural onsets.

Parameter	Voiced	Unvoiced
V/UV	1	1
Pitch	7	0
Linear Phase Coefs	7	0
Cut-off Frequency	2	0
Phase Residuals	34	50
LP Gain	5	5
LSF VQ Index	24	24
Total	80	80

Table 1. Fixed rate bit allocation (per frame).

3. PACKET LOSS CONCEALMENT

The basic principle in the packet loss concealment method is to stretch the packets on each side of the missing packet interval, as illustrated in Figure 1. In this figure, S is the synthesis frame length when no packets are lost, which due to the 50% overlap is equal to half the analysis frame length. Δ_p and Δ_a are the additional lengths of the playout frames prior to and after the packet loss(es), respectively. We see that $\Delta_p + \Delta_a = K \cdot S$ where K is the number of consecutive packet losses. Note the difference in the analysis frame index m and synthesis frame index k as a consequence of lost packets not being given a synthesis index.

In the work presented here, we used $\Delta_a = \Delta_p$ but this could easily be relaxed. For example, if the packet after the loss interval is not yet present in the jitter buffer one could pick a large value for Δ_p and start playout of this packet and then calculate Δ_a when a packet arrives. Furthermore, if both packets are known it might be perceptually better to stretch one more than the other depending on the contents of the packets.

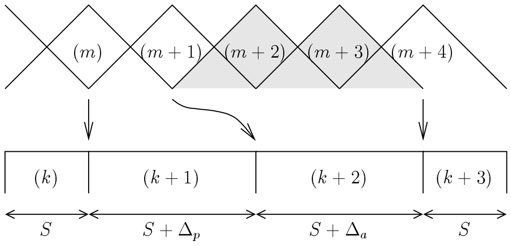

Fig. 1. Principle for packet loss concealment scheme. Shaded frames symbolize packet losses.

As indicated in Figure 1 the stretching of packets is carried out by modifying the point of time in which the amplitudes and frequencies of each packet occurs. This time-scale modification is carried out through a mix of parameter interpolation and overlap-add (OLA). Specifically, the l'th harmonic sinusoidal component is classified for interpolation or OLA by comparison to the corresponding harmonic from the previous synthesis frame. A component in the k'th frame is classified for interpolation if the following three conditions are met ($\hat{a}^{(k)}$ denotes the decoded model parameter a in the k'th frame):

- Both frequencies are below the voicing cut-off frequency of their respective frames, $l\hat{\omega}_0^{(k)} < \hat{\omega}_c^{(k)}$and $l\hat{\omega}_0^{(k-1)} < \hat{\omega}_c^{(k-1)}$.
- The frequency difference is below 70 Hz, $|l\hat{f}_0^{(k)} - l\hat{f}_0^{(k-1)}| < 70$ Hz
- The amplitude ratio is below 5, $\max\left\{\frac{\hat{A}_l^{(k)}}{\hat{A}_l^{(k-1)}}, \frac{\hat{A}_l^{(k-1)}}{\hat{A}_l^{(k)}}\right\} < 5$

The first criterion means that unvoiced components will be overlap-added, whereas the other two prevent interpolation of dissimilar components. Note that unvoiced frames will be synthesized by OLA only.

3.1. Parameter interpolation

For components matched by the three conditions above amplitudes are simply interpolated linearly over each synthesis frame, i.e. for $n = 0 \ldots S^{(k)} - 1$:

$$\tilde{A}_l^{(k)}(n) = \hat{A}_l^{(k-1)} + \frac{\hat{A}_l^{(k)} - \hat{A}_l^{(k-1)}}{S^{(k)}} n \tag{2}$$

Here $S^{(k)}$ denotes the length of the k'th synthesis frame. Likewise, frequencies evolve linearly over the frame, i.e. for $t \in [0, S^{(k)}]$:

$$\tilde{\omega}_l^{(k)}(t) = l\hat{\omega}_0^{(k-1)} + \frac{l\hat{\omega}_0^{(k)} - l\hat{\omega}_0^{(k-1)}}{S^{(k)}} t \tag{3}$$

From this continuous frequency model we determine the discrete phase function:

$$\begin{aligned} \tilde{\theta}_l^{(k)}(n) &= \tilde{\theta}_l^{(k)}(0) + \int_0^n \tilde{\omega}_l^{(k)}(t)dt \\ &= \tilde{\theta}_l^{(k)}(0) + l\hat{\omega}_0^{(k-1)} n + l\alpha^{(k)} n^2 \end{aligned} \tag{4}$$

where $\alpha_l^{(k)} = \frac{1}{2}\left(\hat{\omega}_0^{(k)} - \hat{\omega}_0^{(k-1)}\right)/S^{(k)}$.

In order to avoid discontinuities at frame boundaries the initial phase $\tilde{\theta}_l^{(k)}(0)$ is the final phase of the same component in the previous frame:

$$\begin{aligned} &\tilde{\theta}_l^{(k)}(0) = \tilde{\theta}_l^{(k-1)}\left(S^{(k-1)}\right) \\ &= \tilde{\theta}_l^{(k-1)}(0) + l\hat{\omega}_0^{(k-1)} S^{(k-1)} + l\alpha^{(k-1)}\left(S^{(k-1)}\right)^2 \end{aligned} \tag{5}$$

That is, the initial phase is determined recursively from the pitch of previous synthesis frames back to the frame where the interpolation track was started.

3.2. Overlap-add synthesis

The remaining, unmatched components are synthesized by OLA simply by stretching the overlap region of the analysis frames as sketched in Figure 2. However, after a packet loss the initial phases should be modified in order to compensate for the time offset Δ_a introduced here. Specifically:

$$\tilde{\phi}_l^{(k)} = \hat{\phi}_l^{(k)} - \Delta_a l\hat{\omega}_0^{(k)} \tag{6}$$

This step ensures that overlap-added components are properly matched to components synthesized by interpolation.

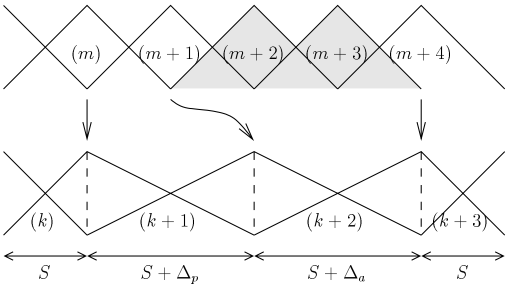

Fig. 2. Overlap-add synthesis in case of packet loss. Shaded frames symbolize packet losses.

4. EXPERIMENTAL RESULTS

In Figure 3 an example waveform resulting from the proposed method is shown in case of packet losses. We see that the structure of the missing parts is well synthesized.

Simple listening tests have been carried out to investigate the performance of the method employed. The tests were conducted using a five point degradation score (Degradation Category Rating): degradation inaudible 5, audible but not annoying 4, slightly annoying 3, annoying 2, and very annoying 1 (see [13]). 12 untrained listeners participated. The test subjects were asked to grade the degradation of the signals relative to the original. Two test signals were used with each consisting of one speaker uttering one sentence. Three different realizations of four different cases of random packet losses were graded.

In table 2 the results of the listening tests are shown in the form of a mean score and a standard deviation for each test case. It can be seen that the average degradation due to the coding process has been graded a little below 4 (audible but not annoying). The effectiveness of the proposed packet loss concealment strategy is evident in that both the 10% and 20% packet loss cases are graded above 3 (slightly annoying), whereas the degradation in the 30% cases is more distinct and thus have received lower scores. These tests

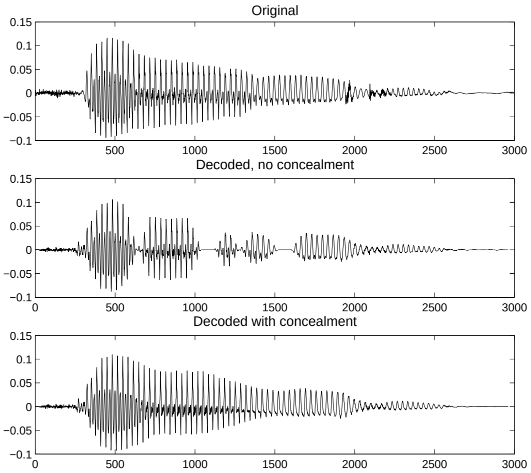

Fig. 3. Example of packet loss concealment. The "Decoded, no concealment" sequence is obtained by silence substitution in lost frames.

Packet loss	Mean Score	Std. Dev.
0%	3.8	0.9
10%	3.3	0.8
20%	3.2	0.9
30%	2.6	0.7

Table 2. Results of listening tests (mean score and standard deviation).

show that at average packet loss concealment can be successfully conducted in the compressed domain using the proposed methods. In fact, the degradation is only about 0.5 for packet losses of $10 - 20\%$ relative to the coded speech. The degradation is generally perceived as the synthesized speech becoming increasingly more tonal for higher packet losses. Also, the coded signal is slightly more tonal than the original.

5. CONCLUSION

In this paper a method for compressed domain packet loss concealment along with a sinusoidal speech coder for packet switched networks have been presented. The method has been evaluated by means of listening tests indicating that it reduces the consequences of packet losses with respect to perceived quality greatly. We therefore conclude that the combination of a sinusoidal coder and packet loss concealment operating on the compressed parameters provides an appealing solution for packet switched networks.

6. REFERENCES

[1] D. J. Goodman, G. B. Lockhart, O. J. Wasem, and W. C. Wang, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Trans. ASSP*, vol. 34(6), pp. 1440–1448, 1986.

[2] J. Lindblom, *Packetized Speech Transmission - Combatting the Packet Loss Problem*, Lic. thesis, Information Theory Group, Chalmers University of Technology, 2001.

[3] J. Lindblom and P. Hedelin, "Packet Loss Concealment Based on Sinusoidal Modeling," in *IEEE Proc. Workshop on Speech Coding*, 2002, pp. 65–67.

[4] Y. J. Liang, N. Färber, and B. Girod, "Adaptive Playout Scheduling Using Time-Scale Modification in Packet Voice Communications," in *IEEE Proc. ICASSP*, 2001, vol. 3, pp. 1445–1448.

[5] F. Liu, J. Kim, and C.-C. J. Kuo, "Adaptive delay concealment for Internet voice applications with packet based time-scale modification," in *IEEE Proc. ICASSP*, 2001, vol. 3, pp. 1461–1464.

[6] S. V. Andersen, W. B. Kleijn, R. Hagen, J. Linden, M. N. Murthi, and J. Skoglund, "ILBC - A Linear Predictive Coder with Robustness to Packet Losses," in *IEEE Proc. Workshop on Speech Coding*, 2002, pp. 23–25.

[7] M. G. Christensen, C. Albøge, S. H. Jensen, and C. A. Rødbro, "A Harmonic Exponential Sinusoidal Coder," in *NORSIG Proc.*, 2002.

[8] C. A. Rødbro and S. H. Jensen, "Time-scaling of Sinusoids for Intelligent Jitter Buffer in Packet Based Telephony," in *IEEE Proc. Workshop on Speech Coding*, 2002, pp. 71–73.

[9] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," in *Journal of ASA*, Apr. 2002, vol. 111(4).

[10] S. O. Rice, "Mathematical Analysis of Random Noise," in *The Bell Systems Technical Journal*, 1944, vol. 3, pp. 282–332.

[11] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," in *IEEE Trans. on Signal Processing*, 1991, vol. 39, pp. 411–423.

[12] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," in *IEEE Trans. on Speech and Audio Processing*, 1993, vol. 1, pp. 3–14.

[13] *Revised Recommendation P.800 (Methods for Subjective Determination of Transmission Quality)*, ITU, Jan. 1996, COM 12-65-E.