ERROR PROTECTION AND PACKET LOSS CONCEALMENT BASED ON A SIGNAL MATCHED SINUSOIDAL VOCODER

Jonas Lindblom and Per Hedelin

Chalmers University of Technology
Information Theory Lab, School of Electrical Engineering,
S-412 96 Göteborg, Sweden.
{jonas.lindblom, per.hedelin}@elmagn.chalmers.se

ABSTRACT

We are proposing an error protection and packet loss concealment system, based on a modified sinusoidal vocoder. By a few additional phase parameters, the vocoder is able to partially track the original signal during voiced segments. This signal tracking allows us to seamlessly switch between the vocoded signal and a primary encoded signal.

A packetized speech transmission system, using PCM data as primary encoding, and the modified vocoder as a sub-coder, is simulated and evaluated using subjective preference testing. Packet loss concealment based on extrapolation of the vocoder parameters is found to perform well, especially at high error rates.

1. INTRODUCTION

When speech data is transmitted over a packet network, such as today's Internet, or future wireless systems, packets containing speech data may be delayed or even lost due to a number of reasons [1]. The problem can be tackled by different strategies. One may focus on the receiver end, and synthesize missing segments based on the previously received signal only [2, 3]. Such methods are conceptually appealing since they can be added to existing systems by modifying only the receivers. They can also be implemented with no, or very low, extra algorithmic delay to the overall system. Receiver based concealment is, however, limited in the sense that it implicitly relies on an assumption that the signal segment that to be recovered is in steady-state. Systems with built in error protection, can potentially do better, but usually at the cost of more bandwidth requirements and delay. In this work, we suggest the use of a signal matched sinusoidal vocoder as error-protection. Our vocoder is signal matched in the sense that it does a decent waveform match to the original signal, making it possible to switch seamlessly between a higher quality, waveform encoded signal, and the vocoded signal, at the receiver.

The use of low-rate sub-codecs for error protection is studied in e.g. [4–6], and is found to work reasonably well, at least at moderate loss rates. The *signal matched sinusoidal vocoder* (SMSV) suggested here, is suitable in such a scenario for a number of reasons: 1) SMSV can be encoded at a low bit rate, and depending on the primary waveform coding technique, adds little overhead in no-loss conditions; 2) SMSV has a good stand-alone quality and, if congestion control is implemented in the network, the primary description may be dropped, and the decoder is still able to reproduce the vocoded signal; 3) Due to the way packets are assembled, the system is particularly robust to frame-erasures. Moreover, the sinusoidal speech model has recently found use for error concealment. The sinusoidal extrapolation scheme in [3], and the jitter buffer techniques in [7], can be applied directly to the SMSV parameters in the receiver.

The paper is organized as follows. In the next Section, a tentative packet structure is described. Then, the signal model for the sinusoidal vocoder is presented in Section 3. The encoder side of the vocoder is outlined in Section 4 and the decoder side in Section 5. In Section 6, we describe our experiments and present the results of some subjective testing that we have performed. The paper is then concluded with Section 7.

2. PACKET STRUCTURE

Initial testing indicates that some of the parameters in the SMSV work best at an update rate of 100 Hz. Therefore, in a telephony application, we assume a frame size corresponding to 10 ms at a sampling frequency of 8 kHz. For our simulations, the primary encoded data corresponds to simple PCM data, and hence 80 such samples are collected in each packet for transmission over the network. In each packet, j, are also two SMSV descriptions, $\boldsymbol{\Theta}(t_j)$ and $\boldsymbol{\Theta}(t_{j+1})$, corresponding to times t_j and t_{j+1} (see Figure 1). This means that the SMSV parameters $\boldsymbol{\Theta}(t_j)$, are transmitted *both*

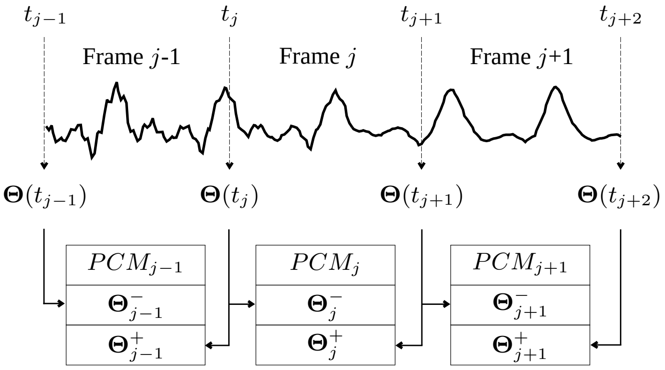

t_{j-1}
t_j
t_{j+1}
t_{j+2}
Frame j-1
Frame j
Frame j+1
$\boldsymbol{\Theta}(t_{j-1})$
$\boldsymbol{\Theta}(t_j)$
$\boldsymbol{\Theta}(t_{j+1})$
$\boldsymbol{\Theta}(t_{j+2})$
PCM_{j-1}
$\boldsymbol{\Theta}^-_{j-1}$
$\boldsymbol{\Theta}^+_{j-1}$
PCM_j
$\boldsymbol{\Theta}^-_j$
$\boldsymbol{\Theta}^+_j$
PCM_{j+1}
$\boldsymbol{\Theta}^-_{j+1}$
$\boldsymbol{\Theta}^+_{j+1}$

Fig. 1. Tentative packet structure. Each packet contains PCM data, as well as SMSV parameters corresponding to both frame borders.

This work is supported by the Swedish strategic research program *Personal Computing and Communication* (PCC).

in packet j, and in packet $j-1$. This redundancy results in robust operation during error conditions. A single packet loss can be handled with full vocoder quality, cf. Section 5.1.

3. THE SIGNAL MATCHED SINUSOIDAL MODEL

To some extent, the signal model for SMSV resembles the model described in [8]. The model is based on linear prediction (LP) analysis and involves an LP production filter as well two LP residual sources: a sinusoidal source, $e_v(n)$, for the voiced case, and a noise source, $e_u(n)$, for the unvoiced case. Both $e_v(n)$ and $e_u(n)$ are designed to have unity power, so that the two can be mixed together in L sub-bands using voicing parameters $\{g_i\}_{i=1}^{L}$ according to

$$e_i(n) = \sigma(n)\left(g_i(n)e_v(n) + \sqrt{1-g_i^2(n)}e_u(n)\right), \quad (1)$$

where $\sigma(n)$ is the slowly varying overall gain, and $e_i(n)$ the excitation signal for band i. Cf. Figure 5. The unvoiced source is equal in [8] and in SMSV, whereas the voiced source differs; the harmonic model of [8] is replaced by a quasi-harmonic model in the SMSV.

3.1. The Quasi-Harmonic Source

The voiced source consists of $M(n)$ sine-waves, where all but the first T are harmonically related. The output of the source, $e_v(n)$, is

$$e_v(n) = A(n)\left(\sum_{i=1}^{T}\sin(\phi_i(n)) + \sum_{k=T+1}^{M(n)}\sin(k\phi_0(n))\right) \quad (2)$$

where $A(n)$ is selected so that $e_v(n)$ has unity power. The instantaneous harmonic phase, $\phi_0(n)$, evolves as

$$\phi_0(n) = \phi_0(n-1) + 2\pi f_0(n)/f_s \quad (3)$$

where $f_0(n)$ is a slowly varying pitch frequency. $M(n)$ is determined by the pitch, $M(n) = \lfloor f_s/2/f_0(n)\rfloor$. The first T sines are "tracked", and their corresponding instantaneous phases evolve according to

$$\phi_i(n) = \phi_i + (\phi_i^+ - \phi_i)\frac{n}{N} \qquad i \in [1,T] \quad (4)$$

where ϕ_i is the starting phase of the current frame, and ϕ_i^+ is the starting phase in the next frame. N is the frame length.

3.2. The Noise Source

The output of the unvoiced source is denoted $e_u(n)$, and consists of white Gaussian noise, with unity power (variance).

4. THE SMSV ENCODER

At the encoder, the parameters of the SMSV are extracted at an update rate of 100 Hz. The pitch and voicing parameters are crucial for natural speech, and we perform careful analysis based on a perceptually weighted LP residual. We rely completely on time domain techniques, for details see [8]. In order to obtain a smooth evolution of the pitch estimate, a 60 ms buffer is used, implying an algorithmic delay of 35 ms on the encoder side.

The LP coefficients are obtained using the autocorrelation method on a 20 ms buffer.

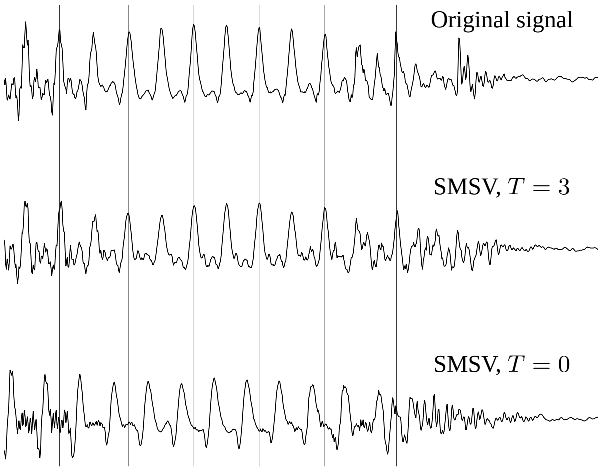


Fig. 2. Illustrating the SMSV's ability to track the original waveform (top) using T=3 tracked sinusoids (middle). Bottom: the SMSV can be used in "normal", unsynchronized, vocoder mode with T=0.

4.1. Signal Matching

Typically, a vocoder does not attempt to mimic the phase properties of the original signal. Time-domain SNR is often negative and does not correspond to subjective speech quality.

The SMSV attempts to do some waveform matching, by tracking the first T quasi-harmonics of the original signal. This is done by fitting a constant amplitude sinusoidal model to 220 samples of the LP residual $r(n)$,

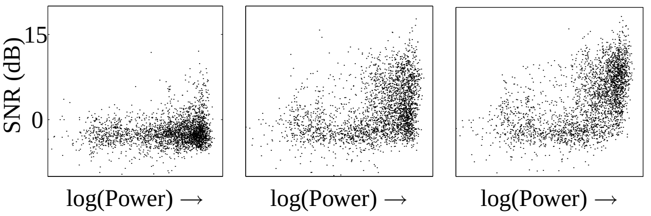
$$\hat{r}_T(n) = \sum_{i=1}^{T} A_i \sin\left(\frac{2\pi\hat{f}_i}{f_s}n + \psi_i\right). \quad (5)$$

When the pitch frequency, f_0, has been determined, the DFT of $r(n)$ is calculated. The frequencies corresponding to the peaks closest to each of the first T harmonics of the estimated pitch frequency, are then determined from the DFT magnitude spectrum. If a peak is too far away from its corresponding harmonic, or to be specific, if $|\hat{f}_i - if_0| < 0.15$, we set $\hat{f}_i = if_0$.

The amplitudes and phases from (5) are then determined by considering a weighted least-squares criterion. Given $\{\hat{f}_i\}_{i=1}^{T}$, the parameters can be solved for explicitly in a recursive fashion as suggested in e.g. [3]. The amplitudes are only used as nuisance parameters, and are not transmitted. From $\{\psi_i\}_{i=1}^{T}$ and $\{\hat{f}_i\}_{i=1}^{T}$ of the current analysis frame, and the corresponding sets from the next frame, we then calculate the starting phase and a suitable $m2\pi$ interval for ϕ_i^+, where m is an integer. Cf. (4).

By letting the first T sinusoids be tracked, we obtain a quasi-harmonic signal, which can be synthesized using smooth, interpolated parameter tracks, and matches the original signal especially during strongly voiced segments. See Figure 2 for a typical example where T=3 sines are tracked. With T=0, the SMSV works in "normal" vocoder mode, i.e. without extra phase information. This is the case for the waveform at the bottom of Figure 2.

By considering the scatter plots in Figure 3, we get an indication of how the signal matching performs. In Figure 3(a), T=0,



(a) T=0 (b) T=3 (c) T=15

Fig. 3. Scatter plots of segmental SNR vs. the logarithm of the frame power (increasing toward the right) using an increasing number of tracked sinusoids, T.

i.e. no tracked sinusoids. The SNR is mainly negative, i.e. the synthesized signal is out of phase. In subfigure (b), T=3 sines are tracked, and a fair share of the high-power frames has positive SNR. In Figure 3(c) is the result for T=15. Now most high-power frames have positive SNR.

One should however not attempt to translate the indicated increase in SNR directly into improved subjective quality. Informal listening tests suggest that by tracking many sines (more than say 10), improvements are obtained. But then a large number of parameters is used, and the improvements are hardly motivated by the increased bandwidth requirements. However, by tracking just a few sines, we get an adequate waveform match for switching between the original signal and the decoded SMSV signal. This is exactly what we utilize in the error protection and packet loss concealment system proposed here.

4.2. Transmitted Parameters

The transmitted SMSV parameters are: pitch f_0, 12th order LP coefficients $\boldsymbol{a}$, overall gain σ, voicing parameters $\boldsymbol{g} = \{g_i\}_{i=1}^{L}$, and the phase parameters for the tracked sines, $\boldsymbol{\phi} = \{\phi_k\}_{k=1}^{T}$. Note the redundancy; the SMSV parameters corresponding to time t_j, are transmitted in both packet j and $j-1$. The phase parameters are unwrapped, so that the parameters transmitted in $\boldsymbol{\Theta}_j^-$, are always in the interval $[0, 2\pi)$. See Figure 1 and 4 for details on how packets are assembled.

$\boldsymbol{\Theta}_j^-$	$f_{0,j}$	σ_j	$\boldsymbol{g}_j$	$\boldsymbol{a}_j$	$\boldsymbol{\phi}_j$
$\boldsymbol{\Theta}_j^+$	$f_{0,j+1}$	σ_{j+1}	$\boldsymbol{g}_{j+1}$	$\boldsymbol{a}_{j+1}$	$\boldsymbol{\phi}_{j+1}$

Fig. 4. Transmitted SMSV parameters.

In this study we are primarily focused on the phase tracking, the switching between the primary coder and the vocoder, and extrapolation of the vocoder parameters during error bursts. Therefore, we do not go into a detailed discussion on parameters quantization. But as pairs of vocoder parameter sets are encoded into each packet (cf. Section 2), it is our belief that state-of-the-art vector quantizers can bring the vocoder bit rate down well below 10 kbps.

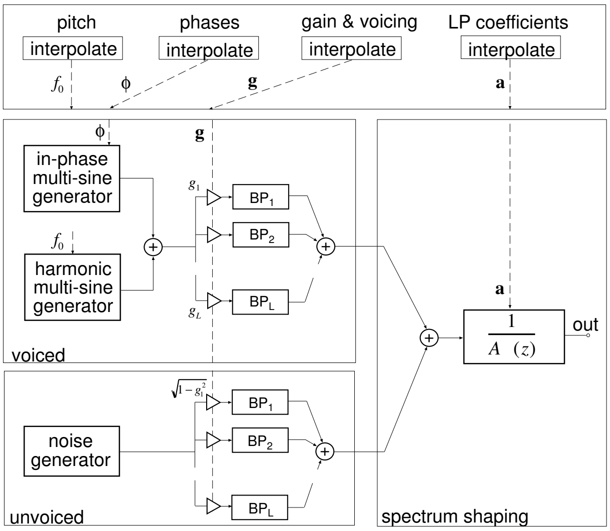


Fig. 5. Principal block diagram of the SMSV decoder.

5. THE SMSV DECODER

All parameters are interpolated on a per sample basis in order to obtain smooth parameter tracks. The LP coefficients, represented by line spectral pairs, are interpolated linearly, whereas the overall gain $\sigma(n)$ and the voicing parameters $\{g_i\}_{i=1}^{L}$ are interpolated on a dB scale. This ensures smooth transitions between regions of varying degree of voicing. A block diagram of the SMSV decoder can be found in Figure 5.

5.1. Packet Loss Concealment

The SMSV is inherently robust to frame erasures since it is based on interpolation between slowly varying parameter estimates. And as the parameters corresponding to one analysis instant are transmitted in two consecutive packets, a single frame loss can be handled with full vocoder quality, if we assume that at least one packet is buffered at the receiver prior to playout. For error bursts longer than one frame, the vocoder parameters need to be extrapolated or interpolated, depending on the amount of buffering on the receiver side. In this work, we focus on extrapolation in order to minimize latency. Thus, for more than one consecutive frame loss, the parameters need to be extrapolated. The LP coefficients are bandwidth expanded as in e.g. [3] in order to avoid unnatural sounds during long error bursts. The extrapolation process is illustrated in Figure 6 and 7. The phase tracking property of the SMSV is used at the beginning and at the end of an error burst, in order to align the vocoded signal with the waveform encoded one.

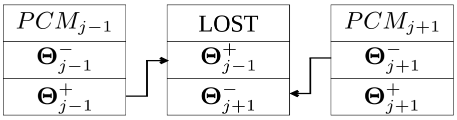

PCM_{j-1}
$\boldsymbol{\Theta}_{j-1}^-$
$\boldsymbol{\Theta}_{j-1}^+$
LOST
$\boldsymbol{\Theta}_{j-1}^+$
$\boldsymbol{\Theta}_{j+1}^-$
PCM_{j+1}
$\boldsymbol{\Theta}_{j+1}^-$
$\boldsymbol{\Theta}_{j+1}^+$

Fig. 6. Concealment of one lost packet using the SMSV.

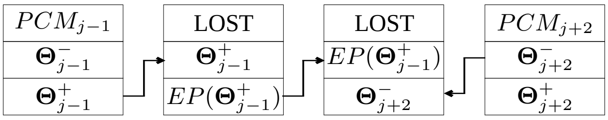

PCM_{j-1}
$\mathbf{\Theta}^-_{j-1}$
$\mathbf{\Theta}^+_{j-1}$
LOST
$\mathbf{\Theta}^+_{j-1}$
$EP(\mathbf{\Theta}^+_{j-1})$
LOST
$EP(\mathbf{\Theta}^+_{j-1})$
$\mathbf{\Theta}^-_{j+2}$
PCM_{j+2}
$\mathbf{\Theta}^-_{j+2}$
$\mathbf{\Theta}^+_{j+2}$

Fig. 7. Concealment of two lost packets using the SMSV. EP denotes extrapolation.

At the beginning of an error burst in say frame j, the LP filter states need to somehow be initialized, or spurious effects may occur. By running the waveform encoded signal from the previous frame through the interpolated LP production filter based on $\boldsymbol{a}^{j-1}$ and $\boldsymbol{a}^{j}$, good results are obtained. At the end of an error burst, the extrapolated LP residual is overlap-added with a residual signal constructed from the first correctly received frame, and the speech signal is synthesized using the corresponding production filter.

6. EXPERIMENTS

In order to test the proposed system, a packetized speech transmission system is simulated. Six sentences from six different speakers, sampled at 8 kHz with 16-bit precision, are processed. Packets containing PCM samples and SMSV parameters with T=3 tracked sines, and $L = 5$ voicing bands, are assembled according to Section 2. Random packet erasures are then generated according to a first-order Markov model. The probability of a frame loss, conditioned on a frame loss in the previous frame, is set to twice the probability of a frame loss conditioned on the previous frame being error free. In this way, we get a slightly bursty channel. Overall loss rates of 10, 20 and 30% are considered.

The processed packets are decoded, and three different packet protection / concealment methods are applied. First of all, we use the SMSV data and the extrapolation methods suggested in Section 5.1. We also evaluate a scenario where the SMSV parameters are always available. This can be useful for networks where congestion control mechanisms sometimes drops the primary encoded data due to network overload, but always lets the sub-coder, i.e. the SMSV through. As a reference method, we also employ our implementation of the G.711 Appendix I packet loss concealment scheme [2].

Pairs of concealed waveforms are presented to 12 test listeners who are forced to indicate their preference. The presentation order of the processed sentences are scrambled for each test listener. Results are presented in Figure 8. We see that the SMSV methods are preferred by the listeners, especially at high error rates. At 10% error rate, all schemes do a fairly good job, and some test listeners noted that it is hard to come up with a clear preference. This is also reflected in the rather large confidence intervals.

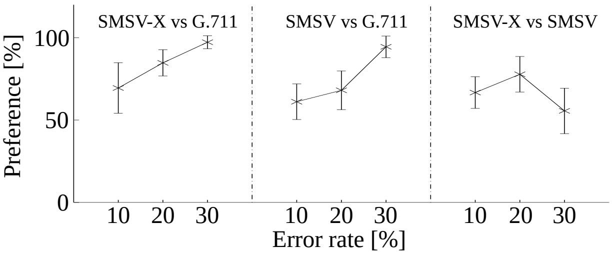


Fig. 8. Results of subjective preference testing. SMSV is the proposed method, SMSV-X is based on the SMSV, which has access to its parameters at all times, and G.711 is the G.711 Appendix I PLC scheme. The vertical indicators are 95% confidence intervals.

7. CONCLUSIONS

A sinusoidal vocoder such as [8] is inherently robust to packet loss, as it synthesizes its output based on slowly evolving (interpolated) parameter tracks. We are proposing a modified sinusoidal vocoder, the SMSV, which is able to track an arbitrary number of sinusoids T in the original signal. The larger the T, the better the waveform match (SNR). The increased SNR does however not correspond directly to increased subjective quality. But by tracking only a few (we use T=3), it is possible to switch between the vocoded signal and a more accurate waveform representation without incurring large discontinuities.

Therefore we suggest the use of the SMSV for error protection and packet loss concealment in a packetized voice transmission system. In a simulated system, we find the suggested methods to work well, especially at high error rates.

The computational complexity of the scheme stems from a number of standard speech processing tasks: estimation of sinusoidal parameters, estimation of pitch frequency, and extraction and interpolation of LP coefficients. Furthermore, it is not necessary to update the LP coefficients at 100 Hz. Preliminary testing indicates that better coding efficiency, without significant loss of subjective quality, is obtained by changing the packetization scheme, and update the LP coefficients at 50 Hz instead.

8. 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] ITU-T Recommendation G.711 Appendix I, *A high quality low-complexity algorithm for packet loss concealment with G.711*, ITU-T, 1999.

[3] J. Lindblom and P. Hedelin, “Packet loss concealment based on sinusoidal extrapolation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, vol. 1, pp. 173–176.

[4] V. Hardman, M.A. Sasse, M. Handley, and A. Watson, “Reliable audio for use over the internet,” in *Proc. INET ’95*, 1995, pp. 171–178.

[5] J. Bolot, S. Fosse-Parisis, and D. Towsley, “Adaptive FEC-based error control for internet telephony,” in *Proc. IEEE Conf. Computer Communications*, 1999, vol. 3, pp. 1453–1460.

[6] T. Morinaga, K. Mano, and T. Kaneko, “The forward-backward recovery sub-codec (FB-RSC) method: A robust form of packet-loss concealment for use in broadband IP networks,” in *Proc. IEEE Workshop on Speech Coding*, 2002, pp. 62–64.

[7] C.A. Rødbro and S.H. Jensen, “Time-scaling of sinusoids for intelligent jitter buffer in packet based telephony,” in *Proc. IEEE Workshop on Speech Coding*, 2002, pp. 71–73.

[8] P. Hedelin, “A sinusoidal LPC vocoder,” in *Proc. IEEE Workshop on Speech Coding*, 2000, pp. 2–4.