PARAMETER INTERPOLATION TO ENHANCE THE FRAME ERASURE ROBUSTNESS OF CELP CODERS IN PACKET NETWORKS

Jian Wang and Jerry D. Gibson

Department of Electrical Engineering
Southern Methodist University
Dallas, TX 75275
e-mail: {wangj, gibson}@seas.smu.edu

ABSTRACT

Frame erasure (FE) robustness is an important quality measure for voice over IP networks (VoIP). Recovery of the erased frames from the received information is crucial to realize this robustness. We allow the lost frames to be recovered from both the "previous" and "next" good frames. We first give quantitative distortion comparisons between predictive and interpolative frame recovery. Then we add FE-robust LSF coding modes to the popular ITU G.723.1 and G.729 CELP coders. These FE-robust modes utilize intraframe LSF VQ and invoke no bit-rate increase for the G.723.1 coder and a small increase (0.4 kb/s) for G.729. Simulations show that FE robust coding with interpolation achieves average spectral distortions 0.7-1.8 dB smaller than that of the original coders. Significant quality improvement was achieved by combined implementation of FE robust coding, LSF and pitch interpolation, and a proposed fixed codebook excitation recovery method.

1. INTRODUCTION

When sending real-time speech packets through IP networks, there is no guarantee of receiving the transmitted packets in a timely manner due to the best-effort nature of the networks. When one or several packets are lost and no effort is made to recover those packets, the perceptual quality of the received speech can deteriorate significantly.

Various schemes can be proposed to alleviate this effect and they are often categorized as encoder-based or decoder-based concealment. Forward Error Concealment(FEC)[1, 2] is popular where redundant speech frames are concatenated with selected packets with a delay. If a frame is lost, the delayed redundant version of that frame may be received correctly to decode that frame.

FEC schemes are effective when the network loss is predictable and extra bandwidth is available. For bandwidth limited applications, decoder-based recovery becomes important. CELP coded speech frames are suitable for this method since many coding parameters show good smoothness between frames. Figure 1 shows a plot of normalized autocorrelation of selected LSF parameters and pitch lag for G.723.1 coded 30ms speech frames. The fact that the autocorrelations are close to 1 shows the smoothness characteristics of the LSF and pitch lag signals.

Some ITU speech coders[3, 4] have built-in mechanisms to process the erased frames based on predictive recovery. These schemes introduce no extra delay because the parameters of the lost frames are recovered from previous good frames. However, both of the above referenced coders quantize LSF parameters via predictive methods. Predictive recovery can cause error propagation to later frames as illustrated in Figure 2. Work in [5] indicates that in VoIP applications, the perceptual quality improvement with interpolation from both the "previous" and "next" correctly received intraframe coded frame is well worth the extra delay introduced.

This research was supported, in part, by NSF grant NCR-9796255.

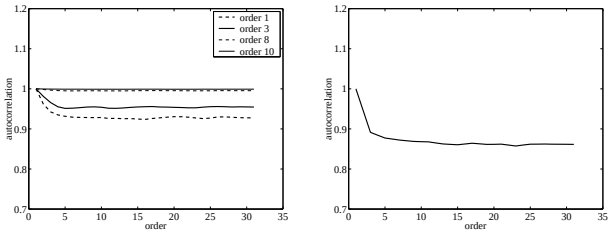

Fig. 1. (a)Normalized autocorrelation of LSF parameters (b) Normalized autocorrelation of pitch lag parameters

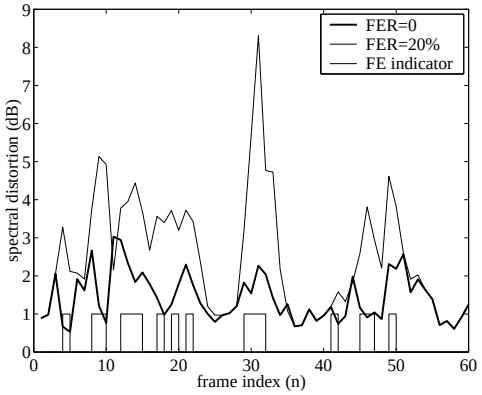

Fig. 2. LSF SD error propagation of G.729 coded speech

In this paper, we first compare the LSF spectral distortion caused by predictive and interpolative FE recovery, then we describe the interpolation method for CELP speech frames. Experi-

mental results are given for ITU G.723.1 and G.729 coders.

2. INTERPOLATION OF LSF PARAMETERS

When coding LSF parameters in CELP coders, square error or weighted square error [6] are usually used to calculate the reconstruction error. On the other hand, the perceptual quality of the LSF quantization is usually measured by spectral distortion (SD). Although it is generally considered that spectral distortion decreases as we quantize LSF more precisely, there are no quantitative relationships between the square error and spectral distortion. To obtain a statistical relationship, we plotted the spectral distortion vs. square error for 10000 30ms frames and found that the average spectral distortion can be approximated as a linear function of the square error. This result is used in section 2.3.

2.1. Expected square error for interpolative recovery

We first calculate the expected square error for an interpolative LSF recovery from intraframe LSF coding. Starting from time $n+1$, L consecutive frames are lost. The interpolation method recovers the lost LSF vector by linear interpolation between the "previous" and "following" good frames. Let the P-dimensional vector, $\mathbf{F}_n = (f_1, f_2, \dots f_P)$ be the n-th frame LSF vector and $\hat{\mathbf{F}}_n$ be the corresponding quantized or interpolated LSF vector; then the interpolated lost LSF vector can be written as

$$\hat{\mathbf{F}}_{n+x} = \frac{L+1-x}{L+1}\hat{\mathbf{F}}_n + \frac{x}{L+1}\hat{\mathbf{F}}_{n+L+1} \tag{1}$$

The LSF parameters can be assumed wide sense stationary. We approximate the quantized LSF vectors by the unquantized version, and taking the expectation of the average square distortion

$$D_L = \frac{1}{L}\sum_{x=1}^{L}\sum_{p=1}^{P}(f_{n+x,p} - \hat{f}_{n+x,p})^2 \tag{2}$$

we can write the expected distortion of these L frames

$$\begin{aligned} ED_{int} = \frac{\Phi(0)}{L}\sum_{x=1}^{L}[1 + \frac{(L+1-x)^2 + x^2}{(L+1)^2} \\ - \frac{2(L+1-x)}{L+1}\phi(x) - \frac{2x}{L+1}\phi(L+1-x) \\ + \frac{2x(L+1-x)}{(L+1)^2}\phi(L+1)] \end{aligned} \tag{3}$$

where $\Phi(.)$ and $\phi(.)$ are the sum of LSF autocorrelations and normalized sum of autocorrelations. They are defined as

$$\begin{aligned} \Phi(\tau) &= \sum_{p=1}^{P} E[f_{n,p}f_{n+\tau,p}] \\ \phi(\tau) &= \frac{\sum_{p=1}^{P} E[f_{n,p}f_{n+\tau,p}]}{\sum_{p=1}^{P} E[f_{n,p}^2]} \end{aligned} \tag{4}$$

2.2. Expected square error for predictive recovery

For predictive recovery, the lost LSF frame is recovered from interframe predictive coded "good" previous received frames by a fixed scalar predictor β and the recovered LSF vector

$$\hat{\mathbf{F}}_{n+x} = \beta^x \hat{\mathbf{F}}_n \tag{5}$$

Note that recovery error could propagate to later frames. This propagation can be forgotten after several "good" frames. For simplicity, we restrict the propagation to one frame in this calculation. Let $\mathbf{e}_n$ be the received residual vector, the effected LSF vector can be written as:

$$\hat{\mathbf{F}}_{n+L+1} = \beta^{L+1}\hat{\mathbf{F}}_n + \mathbf{e}_{n+L+1} \tag{6}$$

The total square error of these L+1 frames will be the sum of the recovered part and propagated part.

Similar to Eq. (3), we can write out the expected distortion for the recovered part as

$$L \times ED_{L,pred} = \Phi(0)\sum_{x=1}^{L}[1 + \beta^{2x} - 2\beta^x\phi(x)] \tag{7}$$

For the propagated part

$$D_{prop} = \sum_{p=1}^{P}(f_{n+L+1,p} - \beta^{L+1}f_{n,p} - e_{n+L+1,p})^2 \tag{8}$$

We take the expectation on both sides. All terms with e_{n+L+1} equal zero since e_{n+L+1} is independent of f_n's and the expectation of e_{n+L+1} equals zero. Ignoring the small e_{n+L+1}^2 term, we get

$$ED_{prop} = \Phi(0)[1 + \beta^{2(L+1)} - 2\beta^{L+1}\phi(L+1)] \tag{9}$$

The expected distortion averaged over $L+1$ frames is

$$\begin{aligned} ED_{Pred} &= \frac{1}{L+1}(L \times ED_{L,pred} + ED_{prop}) \\ &= \frac{\Phi(0)}{L+1}\sum_{x=1}^{L+1}[1 + \beta^{2x} - 2\beta^x\phi(x)] \end{aligned} \tag{10}$$

2.3. Comparison between predictive and interpolative recovery

We calculated the sum of autocorrelations of 100000 30ms LSF as in Table I.

Table I. Sum of autocorrelations of speech LSF parameters(30ms, 10th order)

τ	$\phi(\tau)$	τ	$\phi(\tau)$
0	1	3	0.9941
1	0.9976	4	0.9933
2	0.9956	5	0.9929

From Eq. (3) and Eq. (10) we obtain the optimal predictor $\beta = \phi(1)$ for first order predictions. The ratio ED_{int}/ED_{pred} for $L = 1, 2, 3$ is tabulated in Table II. Note that the expected

average distortion from predictive recovery is greater than that of interpolated recovery by a factor of 2. If we approximate the SD-SE relationship by linear regression on the log-log scale, say $\log(SD) = r\log(sqe) + b$ where r is positive, then

$$\frac{SD_{Pred}}{SD_{int}} = (\frac{ED_{Pred}}{ED_{int}})^r > 1 \quad (11)$$

Table II. Average distortion ratio of predictive and interpolative LSF recovery

L	ED_{pred}	ED_{int}	ED_{pred}/ED_{int}
1	0.1933	0.0734	2.64
2	0.2397	0.1005	2.39
3	0.2743	0.1284	2.14

3. EXCITATION INTERPOLATION

Traditionally, the excitation signals are interpolated based on voice/unvoiced (V/UV) decision on the previous frame and only one of the adaptive or fixed codebook contribution is recovered. The procedure is

i. Get voicing decision on the previous frame;

ii. If previous frame is voiced: set fixed codebook contribution to zero and use previous pitch information, apply attenuated pitch filter to get current excitation,

iii. If previous frame is unvoiced: set adaptive codebook information to zero, use previous gain information, replace excitation signals by a sequence of random numbers normalized by the attenuated gain.

This scheme works for unvoiced frames well. For voiced frames, we observe that there is still some periodicity structure in the fixed codebook excitation signals. Simply replacing the fixed codebook contribution by zeros will not exploit such structure. Further we observe that this structure can be represented by a "pattern" in terms of pulse position and gain which are related to each other. Therefore, we can reconstruct a lost fixed-codebook frame by searching the closest match pattern reflected by its neighboring frames and replace the lost signals by the corresponding frame in the matched pattern. This pattern matching method is similar to [7] except for two differences. One is our search is sub-frame based while [7] is segment based. Secondly, we use pulse gain instead of cross-correlation since pulse signals are scarce.

We applied this pattern matching method to G.729 coded frames and the procedure is illustrated in Figure 3 where fixed codebook subframes 9-10 (320-400 on the samples scale) are lost. We search the closest gain as subframe 8 within 6 subframes and found frame 6 is the closest, so we replace subframe 9 by subframe 6's right neighbor 7. Similarly, we replace subframe 10 by subframe 7's left neighbor, which is 6. The recovered subframes match the original frame except for two pulses.

The pitch lag and gain are always linearly interpolated between the "previous" and "next" good frames.

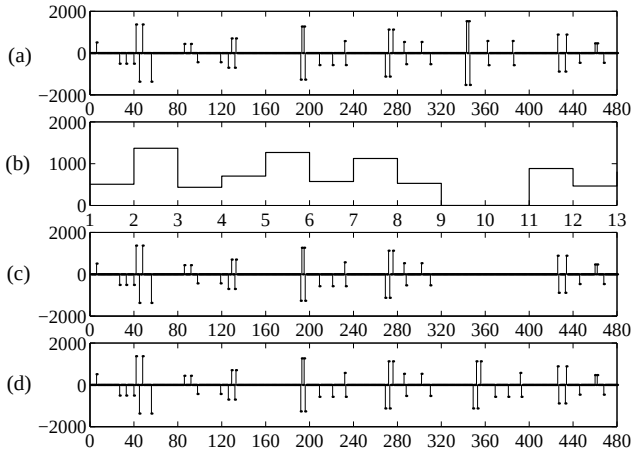

Fig. 3. Interpolation of erased fixed codebook signals by pattern matching. (a) original Fcbk, (b) Fcbk gain, (c) erased Fcbk (d) recovered Fcbk

4. SIMULATIONS

In our simulations, various intra-frame LSF quantization methods were designed for ITU G.723.1 and G.729 speech coders to improve the distortion and FE robustness. We found that split VQ (SVQ)[6] and two-stage VQ-Lattice VQ(VQ-LVQ)[8] achieves smallest spectral distortion. At 24 bits/frame for G.723.1 coding, SVQ and VQ-LVQ achieve similar distortion while VQ-LVQ requires lower complexity. A 22 bits/frame SVQ achieves minimal distortion for G.729 coding, which is 4 bits/frame more than the G.729 predictive LSF coding and yields a 0.4 kb/s rate increase.

We simulate real-time voice over packet networks where each packet contains one frame. Packet loss is approximated by a Markov random process which emphasizes the "bursty" nature of Internet packet loss. Let state "0" stand for a packet being correctly received and "1" be a packet being erased. Let the p be the transition probability from "0" to "1" and q be the probability from "1" to "0" and five loss rates are simulated as listed in Table III.

Table III. Simulated loss rates

rate(%)	p	q
0	0	0
10	.1	.85
20	.2	.70
30	.3	.65
40	.3	.50

The complete recovery process can be summarized here.
On frame erasure,

i. linearly interpolate the LSF parameters from "previous" and "next" good frames;

ii. interpolate pitch lag and gain information;

iii. make V/UV decision based on previous good frame;

iv. if the previous frame is voiced, use pattern matching method to recover the fixed codebook contribution, otherwise;

v. if the previous frame is unvoiced, generate a sequence of random numbers and normalize it by an attenuated version of the previous Fcbk gain.

Figure 4 shows the interpolation performance of the suggested FE-robust LSF quantizer compared to G.729 predictive coding. The SD outliers are important parameters effecting the perceptual quality of the decoded speech and therefore are tabulated in Table IV. Note that the actual packet loss rates we applied to the G.729 predictive coding are smaller by 2% since we allowed extra time-out to indicate a loss event for G.729 so that we can get a fair comparison with interpolative recovery. For example, at 20% loss rate, the actual loss rate we applied to G.729 coding is 18%. This approximation is fairly good since the probability of receiving a packet after its succeeding packet is small.

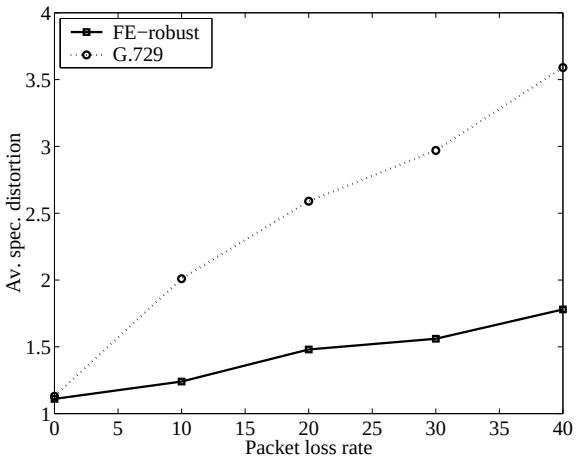

Fig. 4. Average LPC spectral distortion with frame erasure

Note that with 0.4 kb/s rate increase, the added FE-robust LSF coding method achieves 0.7-1.8 dB lower average spectral distortion. The percentage of outliers is also much smaller which yields significant perceptual quality improvement when frame erasures occur.

Table IV. Outliers of LPC spectral distortion with packet loss

frame loss (%)	G.729 Av.SD (dB)	G.729 Outliers(%) $2-4dB$	G.729 Outliers(%) $>4dB$	FE-robust Av.SD (dB)	FE-robust Outliers(%) $2-4dB$	FE-robust Outliers(%) $>4dB$
0	1.13	6	0	1.11	2	0
10	2.01	24	3	1.24	5	0
20	2.59	34	9	1.48	8	2
30	2.97	41	14	1.56	11	2
40	3.59	47	23	1.78	15	3

Informal listening tests show that combined application of FE-robust coding, LSF and Pitch interpolation and fixed codebook recovery on voiced frames achieves significant quality improvement on frame erased speech. Similar simulations were also performed on a G.723.1 coder where a two stage VQ-LVQ was used for FE-robust LSF coding at 24 bits/frame. The results are similar to the Intra-DQ as presented in [5].

The resulting delay from interpolation is the multiplication of the erased frame periods. For example, if three frames are erased in a row, the delay will be $3 * 30ms + RTT/2$, where RTT is the average network round trip delay and ranges from 10-700 ms for a typical network. The maximal acceptable delay for VoIP applications is less than 800ms. Therefore, the delay caused by interpolation may be insignificant compared to the trade-off in speech quality.

5. CONCLUSIONS

By adding FE-robust LSF coding modes and allowing the erased frames to be interpolated from "previous"and "next" good frames, combined with the recovery of the erased fixed codebook signals by the proposed pattern matching method, we introduced an effective method to recover the erased CELP-coded speech frames. Informal listening tests show the quality improvement of this method is significant compared to ITU built-in FE concealment methods. The drawbacks of this method are the extra delay required for interpolation and possibly the extra bit rate for FE-robust coding. Such delay and rate increases may be insignificant compared to other factors effecting the communication channels.

6. REFERENCES

[1] J. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive FEC-Based Error Control for Internet Telephony," *Proceedings - IEEE INFOCOM*, vol. 3, pp. 1453-1460, March 1999.

[2] M. Podolsky, C. Romer and S. McCanne, "Simulation of FEC-based error control for packet audio on the Internet," *Proceedings - IEEE INFOCOM*, vol. 2, pp. 505-515, April 1998.

[3] ITU, *ITU-T G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, ITU 1996.

[4] ITU, *ITU-T G.729: CS-ACELP Speech Coding at 8 kbit/s*, ITU 1998.

[5] J. Wang and J. D. Gibson, "Performance comparison of intraframe and interframe LSF quantization in packet networks," *Proc. 2000 IEEE Workshop on Speech Coding*, Delavan, WI, USA, September 2000.

[6] K. Paliwal and B. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Transactions on Speech and Audio Processing*, vol. 1, No. 1, pp. 3, Jan. 1993.

[7] D. Goodman and G. Lockhart etc, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, No. 6, pp. 1440, Dec. 1986.

[8] J. Pan and T. R. Fischer, "Vector quantization of speech line spectrum pair parameters and reflection coefficients," *IEEE Transactions on Speech and Audio Processing*, vol. 6, No. 2, pp. 106, March 1998.