RECONSTRUCTION OF MISSING PACKETS FOR CELP-BASED SPEECH CODERS

Aamir Husain and Vladimir Cuperman

School of Engineering Science, Simon Fraser University
Burnaby, BC V5A 1S6, Canada
Tel: (604) 291-4371, Fax: (604) 291-4951 email: aamir@cs.sfu.ca, vladimir@cs.sfu.ca

ABSTRACT

A common aspect of speech transmission through packetised networks is the need to consider the discarded (missing) packets as a result of error detection or network overload. The missing packets and the possible mistracking that results in the speech decoder lead to significant quality degradation. In this paper, we introduce a packet recovery technique for CELP based speech coders. The proposed technique extrapolates independently the excitation signal and the short-term synthesis filter. A recovery strategy based on speech classification (voiced, unvoiced, transition, silence) is discussed. The extrapolation of the short-term filter uses a least-squares fading memory polynomial filter applied to reflection coefficients. Objective and subjective quality evaluations of the recovery system applied to the LD-CELP G.728 standard and a variable rate CELP system for random and burst frame erasures are presented. The results indicate that the system is robust up to a frame erasure rate of 10%. Very little degradation in quality was observed at erasure rates up to 3%.

1. INTRODUCTION

The 16 kb/s Low-Delay Code Excited Linear Prediction (LD-CELP) speech codec is likely to be employed in the early phase of Personal Communication Services, until lower rate standards are fully established. This codec has been evaluated using a random errors channel model which is an acceptable method for wired networks. However, speech coders operating in a personal communications environment also need to address bursty errors in order to achieve acceptable performance in applications [1]. A simple realization of a bursty channel model can be obtained by introducing short, medium, and long fades, where fades are characterized as signal fluctuations caused by multi-path, a phenomenon commonly exhibited in portable radio channels. Fading leads to a situation where error detection may be preferable to forward error correction (FEC) and the corresponding channel may be characterized by frame erasures. Frame erasures may also appear in packetized systems due to network overload conditions. A bursty channel model characterized by frame erasures of length 1-6 was introduced in [2].

The study of packet losses in speech coding can be traced back to the work of Jayant *et al* [3] who studied odd-even sample interpolation in PCM and DPCM systems. Goodman *et al* [4] reviewed waveform substitution techniques such as pattern matching in PCM systems. Erdol *et al* [5] presented more recent work in waveform substitution based on interpolative techniques on short-time energy parameters in PCM systems. Leung *et al* [6] considered vector linear prediction in voiced frame reconstruction in CELP coders over frame relay networks.

2. SYSTEM OVERVIEW

In this paper, the environment in which frame reconstruction techniques are tested is based on the LD-CELP 16 kb/s standard (G.728) [7] and a variable rate CELP coder (VAR-CELP) [8]. The packet recovery model (PRM) presented is general in nature and can be applied to other CELP coding mechanisms at rates of 4-16 kb/s.

The VAR-CELP codec uses a modular approach whereby the general structure and the coding algorithm for all rates is based on the structure of the highest bit rate system. Lower bit rates are obtained by disabling codec components. Each speech frame is analyzed by a frame classifier and classified as either voiced, unvoiced, transition, or silence in order to determine the coding rate. The system switches between three distinct codec configurations: 8.0 kbit/s for voiced and transition frames, 4.3 kbit/s for unvoiced frames, and 667 bit/s for silence frames with an overall average bit rate

This work is supported in part by Rockwell International, Digital Communications Division, Newport Beach, CA

0-7803-2431-5/95 $4.00 © 1995 IEEE

of 4.2 kbit/s based on averaging of typical male/female speech files with 40% silence.

The problem of frame losses is more acute in codecs using backward adaptation of the synthesis filter (e.g. LD-CELP) because the choice of filter parameters is dependent on past (possibly incorrect) synthesized speech, resulting in error propagation and inaccuracies in tracking the correct speech signal.

A number of existing reconstruction techniques are based on extrapolating the synthesized speech waveform. In a codec which uses backward prediction, this would imply adaptation on a signal which repeats itself in some deterministic fashion. Our experiments show that such an approach may lead to artifacts in the reconstructed speech. To avoid this situation we introduce a reconstruction procedure in which the excitation and the synthesis filter are extrapolated independently based on class information. The reconstruction of the missing speech packets is then achieved by passing the extrapolated excitation through the updated (extrapolated) short-term synthesis filter.

The LD-CELP standard is already deployed, hence any proposed changes have to address the issue of inter-operability. Best performance in packetised networks would be obtained by changes in both encoder and decoder; however this would lead to complete lack of inter-operability with already deployed codecs. On the other hand, changes in the decoder only, ensure inter-operability but limit the performance improvement under packet losses. We chose a compromise approach in which the only change in the encoder is the addition of class information. This information would be discarded by the decoder using the original standard. The proposed decoder has a default mode in which it can operate without class information at the expense of a performance degradation in packet loss situations.

The reconstruction models used for both the LD-CELP and the VAR-CELP codecs are basically the same. Although the effect of packet losses is significantly different mainly due to forward versus backward adaptation and due to the presence of the adaptive codebook index (pitch lag) for VAR-CELP, we were able to achieve good recovery results for both coders using the same reconstruction model.

A block diagram of the packet recovery model (PRM) is given in Fig. 1. We assume that the codec's output bit-stream is packetized by grouping the information generated by a fixed number of frames into a packet. For example, in LD-CELP, 6 frames of 20 samples each result in a 240-bits packet, while in VAR-CELP, one frame of 160 samples results in a variable-size packet. The speech segment corresponding to a packet will be called a block hereafter. A classifier is introduced at

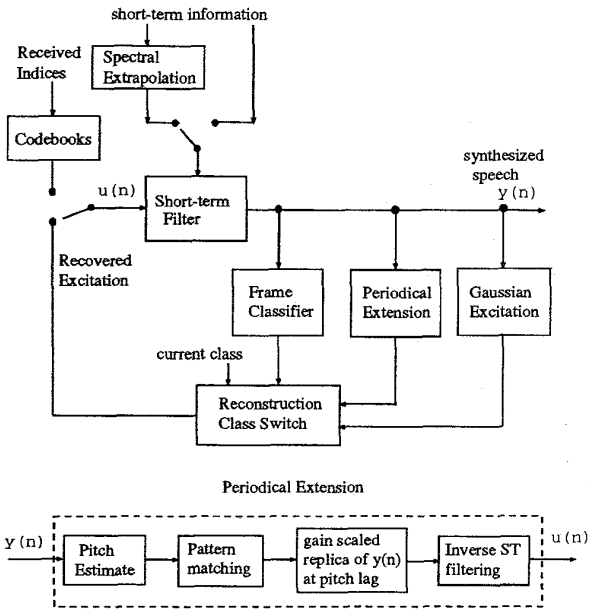

Figure 1: Packet Reconstruction Model

the transmitter and each block is classified as either silence, unvoiced, voiced, or transition. A transition specifies a change from one phoneme to another and includes speech onsets. We assume that the current packet class information is available at the receiver even if the packet is lost. Class information could be transmitted with the previous packet bit-stream, albeit at the expense of additional delay. Alternatively, the class information bits could be protected using a low rate forward error correction code.

In the case of packet(s) loss, the decoder uses a number of different recovery techniques selected as a function of the past and current decoded class. Voiced class excitation recovery is based on periodical extension of the past reconstructed speech and filtering by the inverse of the synthesis filter. Unvoiced class excitation recovery is based on Gaussian modeling. Finally, the past residual excitation memory is recomputed for blocks following lost packets in order to improve the quality at transitions. Details of the excitation recovery model are given in the next section.

3. EXCITATION GENERATION MODEL

The voiced excitation reconstruction model makes use of the previous block pitch estimate, kp_1, as an initial estimate for the current packet. This initial pitch es-

timate is computed at the receiver based on previous reconstructed speech. The estimate is then refined using a pattern matching procedure to select the best lag, kp_2, in a window around the initial estimate [10].

An extrapolation of voiced excitation is obtained by passing through the inverse short-term synthesis filter a scaled replica of previously synthesized speech with a lag of kp_2. The scaling coefficient is obtained by an empirical procedure which combines the gains obtained in the pattern matching procedure together with an estimate obtained from the trajectory of the main pitch-pulse peak.

A Gaussian generated signal was chosen as an appropriate model to approximate actual unvoiced residual signals. The parameters of the Gaussian process for unvoiced excitation extrapolation are estimated from the previous excitation signal [10].

Randomization is used in the excitation extrapolation for offset transitions. For offset frames, a transition from voiced excitation to unvoiced excitation is positioned in the middle of the missing block.

For lost transitions the frame residual excitation memory is corrupted. This leads to inadequate voicing in the following voiced frames as a result of absent or misplaced pitch pulses. The stochastic excitation and adaptive codebook information in the first correctly received block are used to improve the past residual excitation memory before synthesising the correctly received block. This is done by computing an initial residual excitation for the correctly received block and then extrapolating it backward to re-calculate the past residual excitation memory.

4. SPECTRAL EXTRAPOLATION

The spectral extrapolation of the short-term filter is based on the least-squares fading memory polynomial filter, which makes use of the discrete Laguerre polynomials [9]. The extrapolation is applied to reflection coefficient trajectories, rc_n, where rc_n is a vector of the current reflection coefficients at time n, with dimension equal to the order of the short-term filter. The spectral information rc_n is extrapolated by a polynomial in k of degree m, $[p^*(k)]_n$ where k is an index pointing back on the time scale and n is the current time instant and is included to show that the polynomial is found on data up to rc_n. The objective is to minimize,

$$e_n = \sum_{k=0}^{\infty} \{rc_{n-k} - [p^*(k)]_n\}^2 \theta^k \qquad (1)$$

where θ^k is a discount factor, whose value decreases as k increases, provided $\theta < 1$. $[p^*(k)]_n$ may be expressed as a linear combination of the discrete Laguerre polynomials $\varphi_j(k)$,

$$[p^*(k)]_n = \sum_{j=0}^{j=m} (\beta_j)_n \varphi_j(k) \qquad (2)$$

Replacing $[p^*(k)]_n$ in eqn 1 by eqn 2 and solving for minimum e_n with respect to $(\beta_j)_n$ results in the best choice of the coefficients $(\beta_j)_n$ which minimize the error e_n.

5. SIMULATION RESULTS

Our objective was to develop a recovery algorithm which when used in a CELP codec over a noisy channel with packet losses would result in a performance as close as possible to that of the system under error free conditions. Ideally, a difference of the order of less than 0.5 MOS (Mean Opinion Score) would be desirable. The speech decoder should attempt to estimate speech during short and medium fades while during longer fades the reconstructed speech output should decay progressively and recover as quickly as possible after a fade to its error free state. A burst length of one packet corresponds to a loss of 120 samples of speech information in LD-CELP and 160 samples in VAR-CELP. Burst lengths of 1, 3 and 6 packets approximately correspond to short, medium and long bursts respectively.

Table 1 shows the Segmental Signal-to-Noise Ratio Performance of the system under various frame erasure rates (FER) as well as different burst lengths. The results of a Mean Opinion Score (MOS) test are presented in Table 2. The performance of the proposed PRM system is compared with a system in which recovery is based on simple repetition of the indices from the last correctly received packet. This simple recovery system is referred to as the reference system (REF) in Tables 1-2. It should be noted that the reference system based on the simple recovery technique described above performs much better than a system in which the lost packet information is replaced by a random choice of indices. Note that random index substitution may occur in applications if packets affected by fading are transferred to the decoder when frame erasures are not detected.

At frame erasure rates of 3% and a burst length of one packet, MOS evaluation tests have shown only a relatively minor perceptual degradation of the PRM systems over the error free systems. For longer bursts some performance degradation occurs mainly due to mistracking of speech transitions, though the MOS score for the PRM systems are still much better than that of the reference systems. MOS test results show that the

System	clean	FER=3%		FER=10%	
		PRM	REF	PRM	REF
A, bl=1	17.19	13.94	12.75	9.25	6.49
A, bl=3	17.19	15.04	14.51	11.46	9.34
A, bl=6	17.19	16.68	16.02	9.91	8.59
B, bl=1	11.26	9.08	8.64	6.01	5.34
B, bl=3	11.26	7.82	7.42	6.07	5.65
B, bl=6	11.26	9.63	9.40	6.16	5.96

Table 1: Segmental SNR values for FER=3%, 10%, burst length=1,3,6, A) LD-CELP and B) VAR-CELP

	clean	FER=3%				FER=10%	
		bl=1		bl=3		bl=1	
		PRM	REF	PRM	REF	PRM	REF
A	3.85	3.74	3.29	3.45	2.87	3.17	2.66
B	3.70	3.55	3.40	3.25	3.10	3.24	3.03

Table 2: MOS for FER=3%, 10%, burst length=1,3, for systems A) LD-CELP and B) VAR-CELP

LD-CELP-PRM system achieves good recovery performance for short burst lengths at frame erasure rates as high as 10% (MOS is approximately 3.2 at 10% FER). The proposed PRM system achieves results better by 0.4-0.5 on the MOS scale when compared to the reference system.

MOS test results show that the VAR-CELP-PRM system achieves good recovery performance for short burst lengths at frame erasure rates as high as 10%. The difference between the recovered and reference system is not as much as in the case of the reconstruction model tested on the LD-CELP system. A possible reason is the better pitch tracking possible in VAR-CELP due to adaptive codebook information still preserved from the preceding frame in the reference system.

6. CONCLUSIONS

Quality evaluations of the recovery system applied to the LD-CELP G.728 standard and a variable rate CELP system for random and burst frame erasures indicate that the system is robust up to a frame erasure rate of 10%. Very little degradation in quality was observed at erasure rates up to 3%. Further improvements in residual excitation computation for transitions is currently being addressed. Future possible work includes improving short-term filter and predicted gain mis-alignments due to pitch errors in the recovery system applied to the LD-CELP codec. Correcting adaptive codebook misalignments in the recovery system used in the VAR-CELP codec will also be investigated.

7. REFERENCES

[1] S. Dimolitsas, "Standardization of Speech Coding Technology for Network Applications", IEEE Communications Magazine, October 1993.

[2] V. J. Varma, "Testing of 8 kb/s Speech Coders for Usage in Personal Communications Systems", Bellcore Internal Technical Note, 1992.

[3] N. S. Jayant and S. W. Christensen, "Effects of Packet Losses in Waveform Coded Speech and Improvements Due to an Odd-Even Sample-Interpolation Procedure", IEEE Trans on Communications, Vol. COM-29, No 2, Feb 1981.

[4] D. J. Goodman, G. B. Lockhart, O. J. Wassem and W.-C. Wong, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications", IEEE Trans on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No 6, Dec 1986.

[5] N. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of Missing Speech Packets Using the Short-Time Energy and Zero Crossing Measurements", IEEE Trans on Speech and Audio Processing, Vol. 1, No 3, July 1993.

[6] T. W. Leung, W. P. Blanc and S. A. Mahmoud, "Speech Coding over Frame Relay Networks", Proc. IEEE Intl Workshop on Speech Coding, Oct 1993.

[7] J. H. Chen, "High-quality 16kb/s speech coding with a one-way delay less than 2 ms", Proc IEEE Int. Conf. Accoust., Speech, Signal Processing, Albuquerque, New Mexico, pp. 453-456, April 1990.

[8] R. Zopf, M. Cuperman and V. Cuperman, "Real-Time Implementation of a Variable Rate CELP Speech Coder", Proc Int. Conf. on Signal Processing Applications and Technology, Dallas, October 1994.

[9] N. Morrison, "Introduction to Sequential Smoothing and Prediction", McGraw-Hill, New York, 1969.

[10] A. Husain, V. Cuperman, "Classification and Spectral Extrapolation based Packet Reconstruction for Low-Delay Speech Coding", Proc IEEE Globecom Conf, San Francisco, November 1994.