MULTIPLE DESCRIPTION SPEECH CODING FOR ROBUST COMMUNICATION OVER LOSSY PACKET NETWORKS

Wenqing Jiang and Antonio Ortega

Integrated Media System Center
Department of Electrical Engineering-Systems
University of Southern California
Los Angeles, CA 90089-2564
E-mail:[wqjiang,ortega]@sipi.usc.edu

ABSTRACT

Robust speech communication on unreliable channels is one of the key research areas in the development of the voice-over IP (VoIP) technology. In this paper, we propose a multiple description coding (MDC) based speech packetization scheme to combat packet losses. The basic idea is to encode each input speech frame into multiple packets, each of which can be independently decoded. Explicit redundancy is added such that each packet can render an acceptable signal reconstruction of the original frame. Unlike previous appoaches using explicit redundancy for loss recovery [1], we propose to improve the redundancy coding efficiency using context adaptive techniques. Simulation results on independent packet losses show that the proposed scheme gives better average reconstruction audio quality at low loss rates ($\leq 20\%$) compared to that of previous works.

1. INTRODUCTION

Recent years have seen active research in the area of voice-over IP (VoIP) or Internet telephony since packet networks can provide many appealing features that cannot be realized by existing systems, e.g. statistical multiplexing of packets and integration of voice/video and data. One key issue, however, before the wide deployment of the VoIP technology, is how to achieve acceptable audio quality over unreliable channels. For example, in a best-effort network packets can be dropped in heavily loaded segments if some simple congestion control policies are implemented, e.g. random early drop (RED), or they can be overly delayed when they arrive at the receiver due to the congestion and thus become useless for time stringent applications (e.g. teleconferencing). In hostile channels such as a mobile wireless channel, transmitted packets can be completely corrupted by the channel noise. and thus become useless at the receiver. In these cases if the data transmitted in lost packets or corrupted packets is not recovered significant quality degradation can be observed in the received signal [2].

Numerous research efforts have been aiming at providing quality-of-service (QoS) by redesigning the network infrastructure (e.g., RSVP[3]) thus providing bounds on packet losses or avoiding losses altogether. However, in this paper, we study techniques for best-effort networks to enable recovering from packet losses and to mitigate the drop in audio quality. The motivation is that these techniques can complement QoS based transmission (especially if packet losses still occur even if they are bounded), or at least serve as near term solutions before the wide deployment of QoS networks. Our goal is then to design techniques to enable the signal quality to degrade gracefully in the presence of packet losses. A number of such approaches have been proposed in the literature. Jayant [4] proposed a subsample-interpolation technique in which odd and even samples are sent over different packets and the lost packet is recovered as the interpolation results using the received packet. A retransmission scheme is proposed by Karim [5] for mobile radio systems in which corrupted packets are retransmitted. A DPCM diversity system design to combat packet losses is proposed by Ingle et al. [6] using the multiple description coding (MDC) technique [7]. More recently, a robust audio tool (RAT) scheme is proposed by Hardman et al. [1] in which each packet carries explicitly a redundant version of the previous packet for loss recovery. Similar approaches on forward error control (FEC) mechanisms have also been studied by Bolot at al. [8].

In this paper, we propose a new technique based on MDC for robust speech transmission over lossy packet networks. The basic idea of MDC is to send multiple descriptions of the source over the unreliable link, with the hope that at least one of the descriptions can be received correctly so that an acceptable reconstruction of the signal can be achieved. We do not resort to retransmission for error recovery as in [5] because:(i) most multimedia (e.g., speech or images) communications can accept degraded reconstruction if the degradation is less than a certain threshold; and (ii) for multicast applications, it is not efficient to resend data to the whole group if only a few participants suffer from heavy packet losses.

A MDC system using polyphase transform and selective quantization is proposed in our earlier work [9] which has been shown to give excellent results for robust image coding and is also simple for design and implementation compared to previous MDC works [7]. In this system, each polyphase component of the input signal is coded independently using a fine quantizer and packed into one packet. For error protection, each packet also carries a coarsely quantized version of neighboring polyphase components. In case of channel failures, this coarsely quantized data can be used to recover

the lost packets. The approach for loss packet recovery is similar to that used in the RAT system [1], however, we propose in this work to use context adaptive techniques to further improve the coding efficiency for the redundant data. The basic idea of a context-based coding technique is to make use of the knowledge of the neighborhood statistics for the data to be encoded [10, 11]. Since strong correlation, either linear or nonlinear (e.g. structural similarities) usually exists between different polyphase components of a given signal, this correlation can be certainly exploited in our system for better coding efficiency.

In this paper, we propose a context-adaptive MDC system for robust speech packetization. We will show that the proposed system can achieve better performance for independent packet losses as compared to previous works [4, 1, 8]. The rest of this paper is organized as follows. In the next section, we present the proposed context-adaptive MDC system and its application to robust speech coding. In section 3, experimental results on speech coding are provided for random packet losses. The last section concludes our work with possible future extensions.

2. THE PROPOSED MDC SYSTEM

2.1. Basic System

In Figure 1 we show the proposed context adaptive MDC system. The two quantizers Q_1 and Q_2 are respectively the fine (high rate R_0) and coarse quantizers (low rate ρ). The input X is first split into two subsources Y_1 (all even indexed samples) and Y_2 (all odd indexed samples), each of which is then finely quantized using Q_1 and packed into packet P_1 or P_2. For error protection and recovery, each packet also carries a coarsely quantized version of the other component. For example, P_1 consists of $\bar{y}_1 = Q_1(Y_1)$ and and $\hat{y}_2 = Q_2(Y_2)$. The other packet P_2 is produced similarly.

To incorporate the idea of context adaptive coding, the coarse quantizer Q_2 also has as input the dequantized data from Q_1. That is, the redundant data, $\hat{y_2}$ and $\hat{y}_1$ is obtained as the result of quantization and encoding conditioned on the dequantized data, $\bar{y}_1$ and $\bar{y}_2$, respectively. Since polyphase components Y_1 and Y_2 are generated from the same input X, strong correlation is expected to exist between Y_1 and Y_2. This is true for natural speech or image signals and the correlation has been taken advantage of in a number of coding standards (e.g., G.721 for speech and JPEG for images). If the input X is the subband data from a DCT or a wavelet transform output, linear correlation is approximately removed. However, strong structural similarities still exist between polyphase components. One example is that large magnitude coefficients tend to cluster together and so is the case for smaller magnitude coefficients. This feature has been exploited extensively and proven to be very successful in context adaptive codecs developed recently for image coding applications [10, 11]. As a result, more efficient quantization of the redundant information can be achieved.

Over an unreliable channel, packets P_1 and P_2 may not be able to arrive correctly at the receiver. If only one packet is received, then one finely quantized polyphase component and one coarsely quantized polyphase component are used

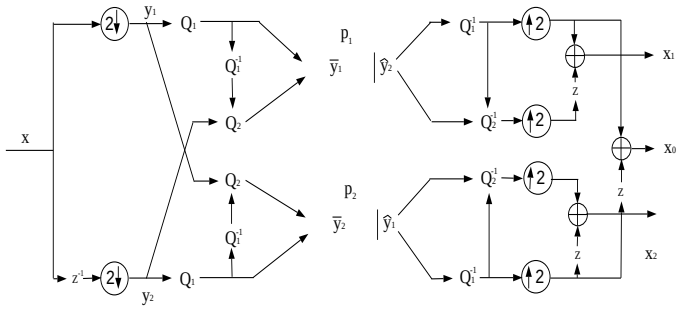

Figure 1: Context-based MDC System

for reconstruction. For example, if only P_1 is received, $\bar{y}_1$ and $\hat{y}_2$ are used for reconstruction. However, if both packets are received, then only finely quantized data $\bar{y}_1$ and $\bar{y}_2$ are used in the signal reconstruction. The system design problem can be formulated as: *Given total bit rate R and packet loss rate p, find the optimal redundancy rate ρ such that the average distortion D is minimized.* Let X be a zero mean random process with variance σ^2. Denote the reconstruction distortion as $D_1 = E|X - X_1|^2$ when only P_1 is received correctly, $D_2 = E|X - X_2|^2$ when only P_2 is received correctly, and $D_0 = E|X - X_0|^2$ when both packets are received. Since these distortions (D_1, D_2, D_0) are all functions of the redundancy rate ρ, the average distortion D is also a function of ρ. Assuming independent packet loss channel with loss rate p, D can be computed as

$$D(\rho) = p^2\sigma^2 + p(1-p)(D_1(\rho) + D_2(\rho)) + (1-p)^2 D_0(\rho)$$

For a given loss rate p and a total bit rate R, the problem of searching for optimal ρ which minimizes $D(R, \rho, p)$, in general, cannot be solved analytically because no closed-form distortion-rate functions exist for a generic random source X. However, analytic solutions do exist for some special cases, for example, when the input X is a i.i.d Gaussian source and high resolution quantization model is used for Q_1 and Q_2 (Interested readers are referred to [9] for more details).

2.2. Speech MDC

An example application of the proposed context-adaptive MDC technique is shown in Figure 2. Each $2N$-sample speech segment is split into two components, Y_1 consisting all even samples and Y_2 of all odd samples. These two components are first finely quantized, e.g., by a PCM or a ADPCM coder, and packed into packets P_1 and P_2 respectively. The dequantized data $\bar{Y}_1$ and $\bar{Y}_2$ are then used to find the prediction residues, $r_1(n)$ and $r_2(n)$ as follows.

$$r_1(n) = y_2(n) - (\bar{y}_1(n) + \bar{y}_1(n+1))/2 \quad (1)$$
$$r_2(n) = y_1(n) - (\bar{y}_2(n-1) + \bar{y}_1(n))/2 \quad (2)$$

These prediction residues are quantized at a lower bit rate using a coarse quantizer Q_2 (e.g. a DPCM coder) and packed into P_1 and P_2 such that $P_1 = \{Q_1(Y_1), Q_2(r_1)\}$ and $P_1 = \{Q_1(Y_2), Q_2(r_2)\}$. Notice that packets, P_1 or P_2, can be decoded independently, i.e., as long as one packet is received, the original $2N$-samples can be reconstructed.

One may notice that, if both packets are lost, then nothing can be recovered for the original $2N$-samples. This constitutes a major drawback of the proposed technique when compared to the RAT scheme, refer to Figure 2, which can always reconstruct at least one packet for consecutive losses [1, 8]. However, as will be shown in the next section, the proposed redundancy coding technique is much more efficient than that used in RAT and the overall performance under independent packet losses still outperforms that of RAT.

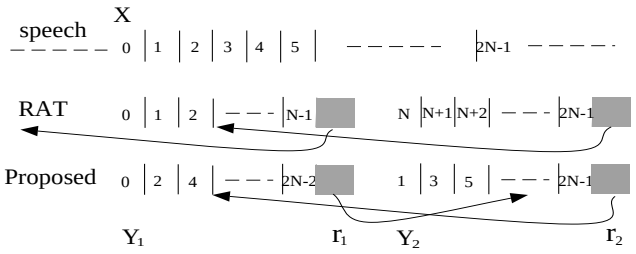

Figure 2: The proposed and the RAT schemes for robust packet speech coding.

3. SIMULATION RESULTS

The speech materials consist of two sentences recorded at 16KHz and 16bits/sample, one male speaker with "A tamed squirrel makes a nice pet" and one female speaker with "Draw every outer line first, then fill in the interior". Each 20ms speech segment is sent in one packet with 320 samples in each packet. A PCM coder is used as the fine quantizer and a ADPCM coder is used as the coarse quantizer in the simulation. With a fixed total rate 16 bits/sample, three different bit allocations are simulated: (A) 13 bits PCM and 3 bits ADPCM; (B) 14 bits PCM and 2 bits ADPCM; and (C) 11 bits PCM and 5 bits ADPCM. The PCM coder is obtained by removing the LSB bits from the original sample, e.g., removing 3 LSB bits to obtain a 13 bits PCM coder. The ADPCM coder is based on the open source coder G.721 and G.723 from SUN Microsystems, Inc..

Three different schemes are implemented and tested in the simulation, namely [1]

1. The subsample-interpolation method by Jayant[4]. The interpolation is the average of two neighboring samples as shown before. Each frame is 16bps PCM coded.
2. The RAT scheme in which each packet carries a x-bit ADPCM coded data of the previous packet for loss protection and the main part is (16-x)bps PCM coded.
3. Every two frames are polyphase transformed and coded by the (16-x)bps PCM coder. The interpolation is the average of two neighboring samples as shown before. The prediction residues are then coded using the x-bit ADPCM. Since the prediction residue is almost uncorrelated, the prediction loop in the ADPCM coder itself is disabled and only the quantization stepsize is adaptively updated based on the input statistics.

To measure the reconstruction quality of speech signal, we use the noise-to-mask ratio (NMR) which measures the relative energy of noise components above the signal's audible masking threshold [12]. The NMR is defined as [12]

$$NMR = \frac{10}{M}\sum_{i=0}^{M-1}\log 10\frac{1}{B}\sum_{b=0}^{B-1}\frac{1}{C_b}\frac{\sum_{k=k_l}^{k=k_h}|D(i,k)|^2}{T_b^2(i)} \quad (3)$$

where M is the total number of frames, B is the number of Critical Bands (CB), C_b is the number of frequency components for CB b, and $|D(i,k)|^2$ is the power spectrum of the noise at frequency bin k and frame i. The k_l, k_h are respectively the low and high frequency bin indices corresponding to CB b.

Table 1: "Squirrel" reconstruction NMR comparison (dB) (A=13:3 B=12:4 C=11:5).

P	*Mean*			*Max*		
	RAT	*MDC*	*JAY*	*RAT*	*MDC*	*JAY*
0.10%A	4.37	-1.41	3.96	10.41	15.25	15.27
0.15%A	6.46	1.37	6.18	15.18	16.70	17.03
0.20%A	8.42	4.78	8.32	15.86	21.51	21.55
0.30%A	11.14	10.10	12.51	20.04	23.23	23.26
0.10%B	2.49	-3.84	3.86	14.05	15.10	15.66
0.15%B	4.38	-0.80	5.35	14.86	15.36	16.44
0.20%B	6.63	3.79	8.53	19.75	18.20	18.19
0.30%B	10.38	10.02	12.47	24.77	21.12	21.17
0.10%C	0.39	-4.28	3.02	15.20	15.04	15.74
0.15%C	3.26	0.74	6.68	17.81	17.34	17.47
0.20%C	5.72	3.89	8.75	18.85	19.24	19.47
0.30%C	9.78	9.99	12.74	33.47	35.31	35.22

In Table 1 and Table 2 we show the reconstruction NMR results for *Squirrel* and *Draw* respectively under independent packet losses at different loss rates (A/B/C stands for different bit allocations as given before). Each loss rate is simulated 100 times and the results are taken as the ensemble averages. The negative NMR values in the table represent cases in which reconstruction noise levels are below the human audio masking thresholds and thus can not be perceived. The data shows that the proposed scheme achieves on an average lowest mean NMR when loss rates are lower than 30%. Actually, the relative performance gain becomes significant at low loss rates (smaller than 20%) as compared to higher loss rates (see Figure 3 for NMR comparisons of the *Draw* sentence). This suggests that the proposed MDC scheme is more suitable for low loss rate applications.

In Table 1 and Table 2 the maximum reconstructed NMR results are also given, which correspond to worst reconstruction scenarios in each simulation. As one can see, the proposed scheme performs worse as compared to the RAT scheme in these extreme cases. As explained before,

[1] In this experiment, to illustrate our ideas, a PCM coder is chosen to encode the primary information and an ADPCM coder is chosen to encode the redundancy. More advanced coders can certainly be used for larger gains, though, more delicate context adaptive techniques are needed.

Table 2: "Draw" reconstruction NMR comparison (dB) (A=13:3 B=12:4 C=11:5).

P	*Mean*			*Max*		
	RAT	*MDC*	*JAY*	*RAT*	*MDC*	*JAY*
0.10%A	1.59	-0.65	6.77	6.22	10.42	12.84
0.15%A	3.61	2.04	8.20	7.89	11.27	13.86
0.20%A	5.72	5.14	10.47	10.67	11.99	17.92
0.30%A	8.52	8.87	13.31	12.20	15.27	18.26
0.10%B	-0.79	-4.05	5.03	6.99	6.10	13.44
0.15%B	1.94	-0.27	7.74	9.97	10.44	17.52
0.20%B	4.13	3.33	9.51	10.01	13.31	18.29
0.30%B	7.58	7.65	12.34	21.04	21.32	21.79
0.10%C	-2.03	-4.64	4.12	5.81	5.77	13.44
0.15%C	0.95	-0.79	7.46	10.54	10.92	16.16
0.20%C	2.70	2.45	9.18	11.05	11.52	17.50
0.30%C	6.68	6.91	12.54	19.40	21.37	21.82

the reason is that when the worst scenario occurs, (i.e., both packets are lost), none of the packets can be recovered using the proposed MDC scheme while RAT can still reconstruct one packet. The higher the loss rate, the more likely this worst scenario will occur, which also contributes to the average performance degradation of the proposed scheme. One possible solution is to introduce more than two descriptions coding, for example, three or four descriptions coding as long as the delay constraints are observed. By doing so, as long as the number of consecutive packet losses is smaller than the number of description packets, the catastrophic case when nothing can be recovered can be avoided.

4. CONCLUSIONS

A MDC speech coding scheme has been proposed in this paper. Simulation results are provided to show that the proposed scheme can achieve better average reconstruction audio quality compared to previous works. Further work is needed to reduce the quality variations among different loss patterns specially at high loss rates.

5. REFERENCES

[1] V. Hardman M. A. Sasse, M. Handley, and A. Watson, "Reliable audio for use over the Internet," in *Proc. INET*, 1995.

[2] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet-loss recovery techniques for streaming audio," *IEEE Network magazine*, Sept./Oct. 1998.

[3] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: A new resource ReSerVation Protocol," *IEEE Network Mag.*, vol. 7, no. 5, pp. 8–18, Sept. 1993.

[4] 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. Communications*, vol. COM-29, no. 2, pp. 101–109, Feb. 1981.

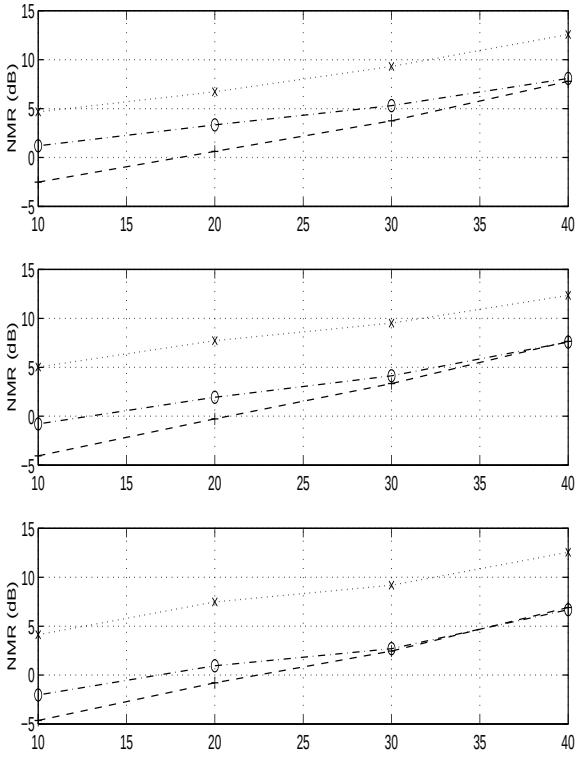

Figure 3: NMR comparisons for the *Draw* sentence. Horizontal axis is the packet loss rate in percentage. Bit allocations: top (A13:3), middle (B12:4) and bottom (C11:5). RAT: o; Proposed: +; JAY: x

.

[5] M. R. Karim, "Packetizing voice for mobile radio," *IEEE Trans. on Communications*, vol. 42, no. 2/3/4, pp. 377–385, Feb./Mar./Apr. 1994.

[6] A. Ingle and V. A. Vaishampayan, "DPCM system design for diversity systems with applications to packetized speech," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 48–58, 1995.

[7] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Information Theory*, vol. 39, no. 3, pp. 821–834, 1993.

[8] J.-C. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive FEC-based error control for Internet telephony," in *Proc. IEEE INFOCOMM'99*, 1999, vol. 3, pp. 1453–1460.

[9] W. Jiang and A. Ortega, "Multiple description coding via polyphase transform and selective quantization," in *Proc. of VCIP'99*, 1999.

[10] C. Chrysafis and A. Ortega, "Efficient context-based entropy coding for lossy wavelet image compression," in *Proc. of DCC'97*, Snowbird, UT, Mar. 1997.

[11] A. Said and W. A. Pearlman, "Low-complexity waveform coding via alphabet and sample-set partitioning," in *Proc. of VCIP'97*, San Jose, CA, Jan. 1997.

[12] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audio noise suppression," *IEEE Trans. on speech and audio processing*, pp. 497–514, 1997.