Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss

Wenyu Jiang and Henning Schulzrinne
Department of Computer Science, Columbia University
1214 Amsterdam Ave, Mail Code 0401
New York, NY 10027, USA
{wenyu,hgs}@cs.columbia.edu

ABSTRACT

Packet loss degrades the perceived quality of voice over IP (VoIP). In addition, packet loss in the Internet tends to come in bursts, which may further degrade audio quality. Using the Gilbert loss model, we infer that changing the packet interval affects loss burstiness, which in turn influences forward error correction (FEC) performance. Next, we perform subjective listening tests based on Mean Opinion Score (MOS) to evaluate the effect of bursty loss on VoIP perceived quality. Then, we compare the perceived quality achieved by two major loss repair methods: FEC and low bit-rate redundancy (LBR). Our MOS test results show that FEC is much preferred over LBR. In addition, our MOS results reveal that, under bursty loss, FEC quality is much better with a moderately large packet interval. Finally, because FEC introduces an extra delay proportional to the packet interval, we present a method of optimizing the packet interval to maximize FEC MOS by considering the delay impairment in ITU's E-model standard.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Performance attributes; H.4.3 [**Communications Applications**]: Computer conferencing, teleconferencing, and videoconferencing

General Terms

Performance, Measurement, Management

Keywords

VoIP; perceived quality; Mean Opinion Score; MOS; forward error correction; low bit-rate redundancy; bursty loss

1. INTRODUCTION

1.1 Packet Loss Repair and Recovery

To recover lost packets in voice over IP (VoIP), the most common scheme is forward error correction, or FEC [20]. It recovers the lost packet in a *bit-exact* form. Figure 1 shows an example of the commonly used Reed-Solomon (RS) code [1]. Its notation is (n, k), where n and k are the number of all and non-FEC data units, respectively. The $\otimes$ symbol is the bit-wise XOR operator used to create FEC (redundant) data. If A is lost, and B is not, by the time C is received, A can be recovered as $B \otimes (A \otimes B)$. *Piggybacking* FEC data in block i onto block $i+1$ reduces the number of packets, and is a commonly used technique. An (n, k) code can recover all losses in the same block *if and only if* at least k out of n packets are received. This introduces a recovery delay of up to $n - 1$ packet intervals. So a longer FEC block means longer recovery time and thus higher end-to-end delay.

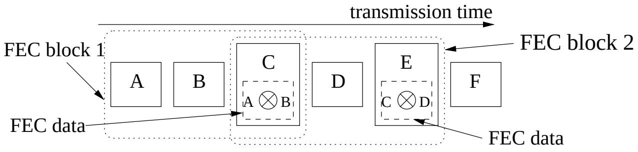


Figure 1: RS (3,2) FEC code with piggybacking

FEC is not necessarily bandwidth inefficient, once we consider the 40 byte RTP [21]/UDP/IP header [1] that has to be present in every multimedia packet. For example, Table 1 illustrates the actual bandwidth of FEC for the 8 kb/s G.729 [11] codec, under different packet intervals (denoted as T).

T	payload	header	no FEC	(3,2) FEC, piggyback
20 ms	20bytes	40bytes	24 kb/s	26.67 kb/s
40 ms	40bytes	40bytes	16 kb/s	18.67 kb/s

Table 1: Bandwidth overhead for G.729 with FEC

An alternative to FEC is low bit-rate redundancy (LBR) [10, 18], which sends a redundant but lower quality version of the same audio. Then a lost main audio packet is approximated by its redundant version.

Finally, if LBR is not used, or if FEC fails to recover a lost packet, the lost audio can be approximated, for instance, by repeating waveforms in the last received packet. This is called packet loss concealment (PLC) [17]. Some codecs such as G.729 [11] and G.723.1 [12] have built-in PLC algorithms, whereas others are defined by the application. A good PLC algorithm can greatly improve perceived quality with no bandwidth overhead.

NOSSDAV'02, May 12-14, 2002, Miami Beach, Florida, USA.
Copyright 2002 ACM 1-58113-512-2/02/0005 ...$5.00.

[1] We are not even counting link layer header bytes yet!

To determine the quality of VoIP under packet loss, the most common metric is the Mean Opinion Score (MOS). In a MOS test, the listeners rate audio clips by a score from 5 to 1, with 5 meaning Excellent, 4 Good, 3 Fair, 2 Poor, and 1 Bad. The resulting average score across listeners is the MOS. The details of MOS test procedures are described in ITU recommendation P.830 [13].

1.2 List of Contributions

Packet loss in the Internet is usually bursty [23, 2], which has certain implications on PLC and FEC. Therefore we have investigated the effect of bursty loss on various aspects of perceived quality in VoIP.

Our main contributions in this paper are as follows. First, we have performed a wide range of MOS listening tests, comparing random versus bursty loss. Then we cover many issues not addressed in previous studies, such as testing different degrees of loss burstiness and packet intervals. Second, we present not only regular MOS test results but also an original study on the MOS quality under FEC and its dependence on loss burstiness and packet interval. Third, our MOS test on FEC and LBR is the first concrete experiment to show that FEC is preferable to LBR. Finally, we have introduced a new method of optimizing FEC MOS through the use of the E-model, considering the delay introduced by FEC.

The remainder of this paper is organized as follows. Section 2 discusses the Gilbert loss model. Section 3 discusses related work. Section 4 gives MOS experiment design rationale and Section 5 presents the corresponding results. Section 6 concludes the paper and lists future work.

2. LOSS MODELING

2.1 The Gilbert Model

The Gilbert model is most commonly used to describe bursty losses, often found in the Internet [23, 2]. It has two parameters, unconditional and conditional loss probability, denoted as p_u and p_c, respectively. The Gilbert model use p_c to quantify loss burstiness.

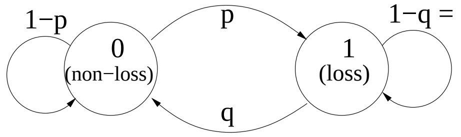


Figure 2: The Gilbert model

An alternative representation is using transition probabilities p and q, as in Figure 2. Apparently $q = 1 - p_c$, whereas p can be computed as:

$$p = \frac{p_u \cdot q}{1 - p_u} = \frac{p_u \cdot (1 - p_c)}{1 - p_u} \qquad (1)$$

Compared to more complex loss models, the Gilbert model allows much easier evaluation in our tests because we need to compare only two parameters instead of many.

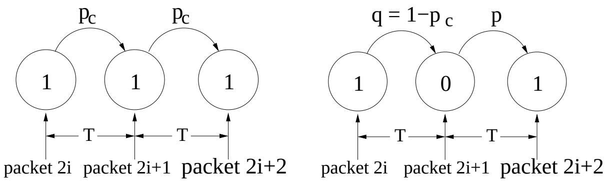
2.2 Loss Burstiness vs. FEC Performance

If a network path is characterized by a Gilbert model with p_u, p_c specified at 20 ms interval, then to correctly simulate the path at 40 ms the simulator should still generate events at 20 ms, but pick either the even or odd sequence, as follows:

```
11000111000001000  original T=20ms event simulation
1 0 0 1 0 0 0 0 0  odd sequence, simulates T=40ms
 1 0 1 1 0 0 1 0   even sequence, also simulates T=40ms
```

Using the above definition, we can derive the new loss pattern when the packet interval changes. In particular, Figure 3 shows how to compute the new conditional loss probability $p_{c,k}$ when the packet interval changes from T to kT, for $k = 2$. The term ${p_c}^2$ and $q \cdot p$ comes from Figure 3(a) and (b), respectively.

$$p_{c,2} = {p_c}^2 + q \cdot p = \frac{(p_c - p_u)^2}{(1 - p_u)} + p_u$$

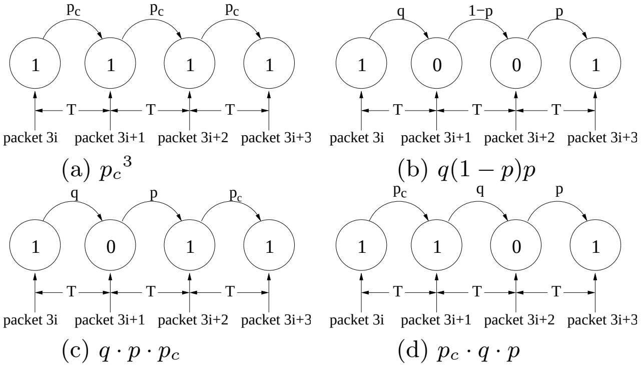


(a) One loss burst (b) Merging of loss bursts

Figure 3: Re-calibrating a Gilbert model when the packet interval increases from T to $2T$

Similarly, Figure 4 illustrates the case for $k = 3$. After adding up individual terms in Figure 4(a)-(d), the new conditional loss probability is:

$$p_{c,3} = {p_c}^3 + q(1-p)p + q \cdot p \cdot p_c + p_c \cdot q \cdot p = \frac{(p_c - p_u)^3}{(1 - p_u)^2} + p_u$$



(a) ${p_c}^3$ (b) $q(1-p)p$

(c) $q \cdot p \cdot p_c$ (d) $p_c \cdot q \cdot p$

Figure 4: Packet interval increases from T to $3T$

The similarity of the above two equations for $k = 2$ and 3 allows a natural generalization to the following formula:

$$p_{c,k} = \frac{(p_c - p_u)^k}{(1 - p_u)^{k-1}} + p_u \qquad (2)$$

Proof of Formula (2) is relatively straight-forward. It is by induction on the term $p_{c,k} - p_u$ using Figure 5. So we do not show the proof here for brevity. We have also confirmed its correctness by Gilbert model simulation.

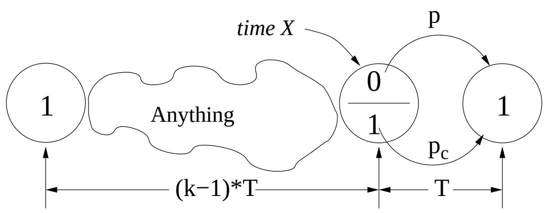
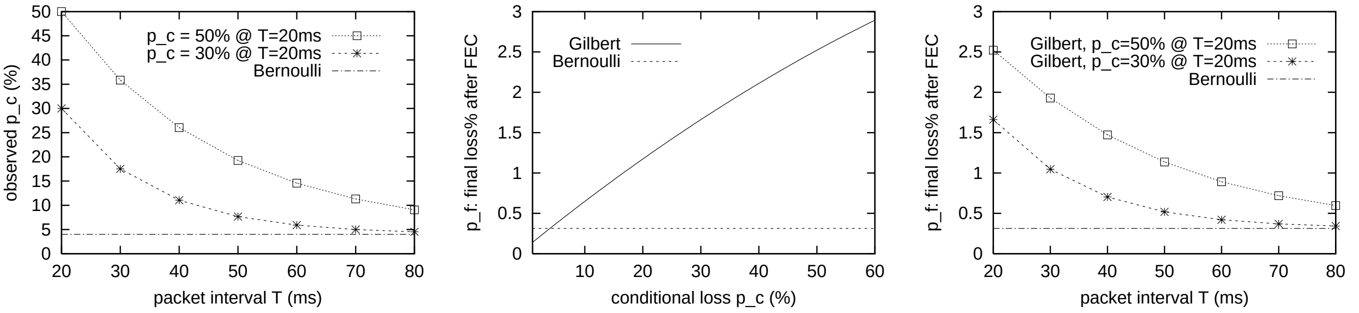


Figure 5: Induction proof for general $T \rightarrow kT$ case



(a) p_c vs. T (b) p_f vs. p_c (c) p_f vs. T

Figure 6: Interaction of packet interval T, loss burstiness p_c, and final loss probability p_f; p_u=4%

Figure 6(a) shows the application of Formula (2). As packet interval T increases, loss burstiness (p_c) decreases. That is, the loss pattern becomes less bursty and more Bernoulli-like. Hence in Figure 6(a) p_c gradually approaches p_u, that is, the random or Bernoulli limit.

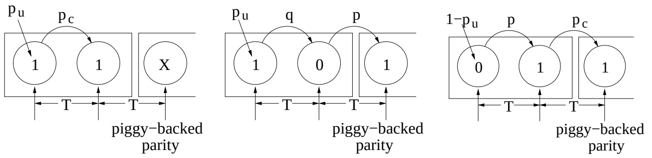
In an (n, k) FEC code, none of the lost packets in a block are recoverable if more than $n-k$ packets are lost. The final loss probability after FEC (denoted as p_f) for an (n, k) code under Bernoulli or random loss [19] is:

$$p_f = p\left(1-\sum_{i=k}^{n-1}\left(\begin{array}{c} n-1 \\ i \end{array}\right)(1-p)^i p^{n-1-i}\right) \tag{3}$$

Frossard [7] gives an in-depth derivation of final loss probability in a Gilbert loss process. However, it assumes no piggybacking. Since piggybacking is widely used, we use Figure 7 to derive the formula for p_f for a (3,2) code. We are investigating whether this procedure is generalizable to any (n, k) code. The formula is as follows:

$$p_f = p_u \cdot (p_c + q \cdot p/2) + (1 - p_u) \cdot p \cdot p_c/2 \tag{4}$$

Applying Formula (4), Figure 6(b) shows that final loss probability increases almost linearly with burstiness p_c. In contrast, the p_f value under Bernoulli loss (horizontal dashed line) is much lower. Therefore FEC is more effective under a non-bursty loss process.



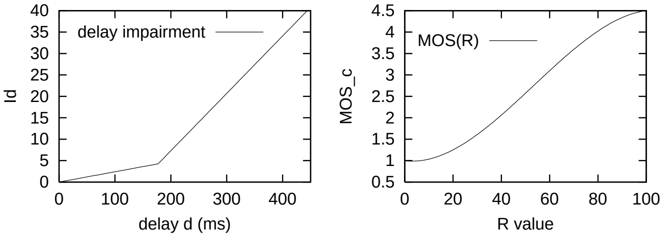
(a) $p_u \cdot p_c$ (b) $p_u \cdot (q \cdot p/2)$ (c) $(1 - p_u) \cdot p \cdot p_c/2$

Figure 7: (3,2) FEC code under Gilbert loss

Because loss burstiness decreases with a larger packet interval, and FEC is more efficient if loss is less bursty, it follows that a larger packet interval leads to better FEC performance, as illustrated by Figure 6(c). Figure 6(c) is very similar to Figure 6(a), except the vertical axis is the final loss probability instead of loss burstiness.

3. RELATED WORK: THE E-MODEL

The E-model [14, 4] is an analytical model for predicting voice quality. It considers various impairment factors including delay, loss, echo, loudness, and frequency response. Each factor is mapped to a score, for instance, loss impairment score I_e or delay impairment score I_d. The mapping from loss probability to I_e is codec dependent, and the E-model provides mappings for some standard codecs. The conversion from delay to I_d is by contrast fixed. Figure 8(a) shows an approximated version [4] of delay to I_d mapping.



(a) delay to I_d mapping (b) R value to MOS mapping

Figure 8: Standard mappings from the E-model

In VoIP, most other factors such as loudness and echo can be simplified to a default number. This leaves only two remaining factors: loss (I_e) and delay (I_d) impairment, from which we can compute a gross score called the R value. Cole and Rosenbluth [4] give the following equation for calculating the R value in VoIP:

$$R = 94.2 - I_d - I_e \tag{5}$$

Finally we apply a fixed mapping [4] from R to MOS, as shown in Figure 8(b). With this mapping, we can derive MOS as a function of loss probability, using standard loss to I_e mappings supplied by the E-model. Figure 9 shows such an example.

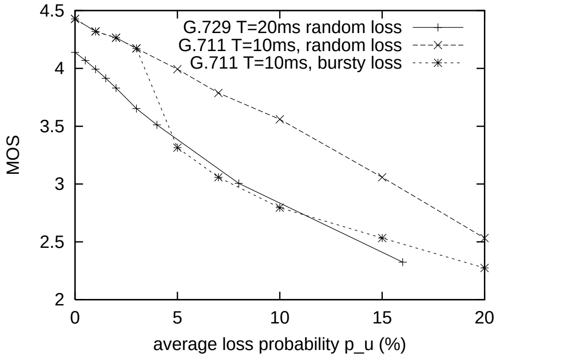


Figure 9: MOS behaviors derived from the E-model

The MOS curves supplied by the E-model have some limitations, however. They either assume a small packet size (e.g., 10 ms/packet for G.711 in Figure 9, which is rare in IP), or that bursty loss mapping is not available (e.g., for G.729), or if available, it does not really specify how "bursty" is bursty (G.711 bursty case in Figure 9). These are the issues we intend to address in our MOS tests.

4. MOS TEST EXPERIMENT DESIGN

4.1 Objectives

The main purposes of our MOS listening tests are threefold: First, to examine how bursty and random loss affect perceived quality differently. Second, to compare the quality of FEC and LBR. Finally, for FEC, how we can maximize its quality.

Based on these objectives, and the observation that bursty loss is more common in the Internet, we have designed our experiments with the following emphases:

- Random vs. bursty (Gilbert) loss model
- Compare FEC and LBR, mostly under Gilbert loss
- MOS with or without FEC under a wide range of loss probabilities (p_u), loss burstiness (p_c) and packet intervals (T).

So we have designed two test sets, $\mathcal{N}_1$ and $\mathcal{N}_2$. We cover the first two items in $\mathcal{N}_1$, and the third item in $\mathcal{N}_2$, as described in Table 2 and 3. In Table 2, every two rows form a comparison pair, for example, Bernoulli (A) vs. Gilbert (B).

4.2 Design of MOS Test Sets

Case ID	Configuration by default $p_c = 30\%$, T=30 ms	loss rate (p_u) 4%	8%	12%
A	Bernoulli	3.92	2.92	2.37
B	Gilbert	3.49	2.71	2.19
C	DoD-LPC LBR	3.23	3.12	2.27
D	FEC(4,3)	3.61	3.83	3.31
E	DoD-CELP LBR	3.21	2.96	3.06
F	FEC(3,2)	3.90	3.96	3.44
G	DoD-CELP LBR Bernoulli	4.06	3.60	3.08
H	FEC(3,2) Bernoulli	4.62	4.33	3.75
I	G.723.1 LBR	3.35	3.05	2.33
J	FEC(2,1)	3.99	3.90	3.44
K	G.723.1 Bernoulli	3.75	3.46	2.83
L	FEC(2,1) Bernoulli	4.61	4.13	3.96
M	AMR12.2+6.7 LBR	3.90	3.55	3.03
N	AMR12.2+FEC(3,2)	4.01	4.03	3.24
O	G.723.1 rat w. state repair	3.17	2.79	2.37
P	G.723.1 rat original	3.04	2.50	2.22

main codec: G.729, except for case M, N (AMR)
secondary codec: G.723.1 runs in 6.3 kb/s
piggybacking: enabled in all FEC codes
packet interval: T=30 ms, except for case M, N (T=20 ms)
loss model: Gilbert with p_c=30% by default

Table 2: Test set $\mathcal{N}_1$: compares PLC, FEC, LBR under bursty and random loss, along with MOS results

The loss range of 4-16% in Table 2 may seem high for the Internet, but we believe it is applicable because the performance of the Internet differs dramatically and is often poor across international paths. Moreover, even a good path may experience temporary congestion, calling for the need to measure its transient state service quality.

Codec	p_u	p_c	T	FEC code
G.729	4-16%	30 or 50%	20-60 ms	(3,2) piggyback

Note: T value of 60 ms is only measured for p_c=50%

Table 3: Test set $\mathcal{N}_2$: compares MOS under different loss burstiness, with or without FEC

4.3 Design of an Optimal LBR mechanism

The original LBR in the RAT program [10] has two flaws:

- Main audio codec decoder state drift
- Inescapable decoder state drift in redundant audio codec

Almost all codecs (except G.711) are frame based and have decoder states. The state drift caused by a single frame loss lasts much longer than one frame [22]. The original LBR specification does not use the redundant audio to help restore or repair main codec's decoder state.

The second problem is due to packet alignment order. In all LBR implementation such as the Robust Audio Tool (RAT) [10], redundant audio packet i is sent at a later time, generally piggybacked onto main audio packet $i + l$, where $l = 1, 2, 3...$ is the *lag*. However, by the time main audio packet i is lost, then redundant audio packet $i - l$ is also lost (assume piggybacking), causing decoder state drift in redundant audio from $i - l$. So the redundant audio for packet i will surely not be decoded correctly, further reducing voice quality.

Therefore, we have designed a new LBR implementation that addresses both issues. The receiver performs decoder state repair by first decoding the redundant audio, then re-encoding it using a duplicate main encoder, and finally decoding it again using the main decoder. Packet alignment is fixed by piggybacking redundant audio corresponding to packet i with main audio packet $i - l$ instead of $i + l$. Finally, alignment should be carefully designed to account for different codec look-ahead delays and frame lengths. For example, G.729 and DoD-CELP [3] have a look-ahead delay of 5 and 15 ms respectively. So we have designed an optimal alignment order in Figure 10 that ensures perfect timing synchronization between G.729 and DoD-CELP frames during playback. In Figure 10, if a G.729 packet i is lost, the receiver should use the DoD-CELP audio decoded from packet $i - 1$ as a replacement.

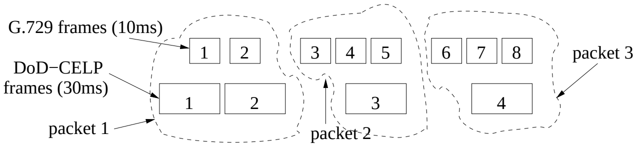
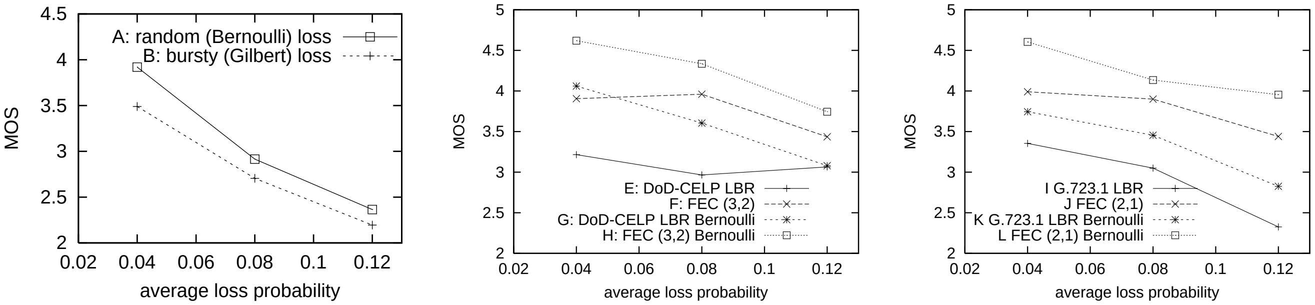


Figure 10: An optimal LBR frame/packet alignment, G.729 + DoD-CELP, lag is 1 packet or 30 ms

4.4 Statistical Properties of MOS Results

We have 20 listeners in our MOS tests. However, the precision of our MOS results is as good as traditional MOS tests with 80 listeners. This is because we ask the listeners to grade at 0.1 granularity, where one can give a score like 3.6 instead of 3 or 4. We expect this method to improve MOS accuracy and also reduce variance. Table 4 confirms



(a) no FEC

(b) FEC and DoD-CELP LBR

(d) FEC and G.723.1 LBR

Figure 11: Random vs. Gilbert loss on MOS quality; G.729 as main codec, T=30 ms, $p_c = 30\%$

that this is indeed true. The standard deviation in our test is only 0.5 MOS on average, whereas traditional MOS tests based on integer scoring typically have a standard deviation around 1.0 MOS [6].

MOS test feature	Our setting	Traditional
No. of listeners, N	20	80
MOS Grading	0.1 granularity	integer scoring
mean standard deviation	≈ 0.5 MOS	≈ 1.0 MOS [6]
90% confidence interval	≈ 0.193 MOS	≈ 0.186 MOS

Table 4: Improved accuracy via 0.1 MOS scoring

Therefore with 20 listeners, we have achieved nearly the same confidence interval (0.193 MOS) as in a traditional integer scoring test with 80 listeners (0.186 MOS), as shown in Table 4. In addition, half of the 20 listeners are from New York and the other half from California. This makes the results much less biased based on the region of people.

5. MOS TEST RESULTS

5.1 Test Set $\mathcal{N}_1$: Random vs. Bursty Loss

Figure 11(a) compares the perceived quality under bursty (curve B, dashed line) and random loss (curve A, solid line). It is obvious that bursty loss has lower MOS. This concurs with the G.711 curves derived from the E-model in Figure 9, where bursty loss also results in lower quality.

Figure 11(b) and (c) compares the quality of LBR and FEC under bursty and random loss. In Figure 11(b), DoD-CELP LBR works much better under random loss (curve G) than bursty loss (curve E) except at 12% loss rate where they have the same MOS. Not surprisingly, its corresponding (3,2) FEC code also has better MOS under random loss (curve H) than under bursty loss (curve F). It is evident in Figure 11(b) that FEC is always better than LBR, whether under bursty or random loss. We will see similar results in section 5.2. Figure 11(c) shows a similar trend, for G.723.1 LBR and its comparison FEC code.

Figure 11 shows that under either FEC or LBR, or even when neither technique is used, audio quality is always better under random loss. However, this only compares between random and bursty loss. Within bursty loss, the degree of burstiness may not have a monotonic effect on quality degradation. That is, higher burstiness sometimes does not lead to lower MOS. This is to be explained in section *5.3.1*.

Finally, Figure 12 shows the optimality of our LBR implementation, which is described in section 4.3. Curve I corresponds to our optimal LBR algorithm. It produces a slightly but consistently higher MOS rating than the RAT-style LBR with decoder state repair but non-optimal alignment (curve O). The gap is even further between curve I and the pure RAT-style LBR (with neither enhancement, curve P). Therefore, our LBR algorithm produces the best MOS among other implementations, and thus lends itself to a fair comparison with FEC.

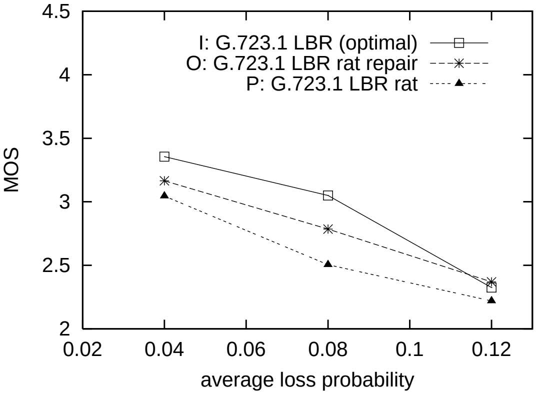


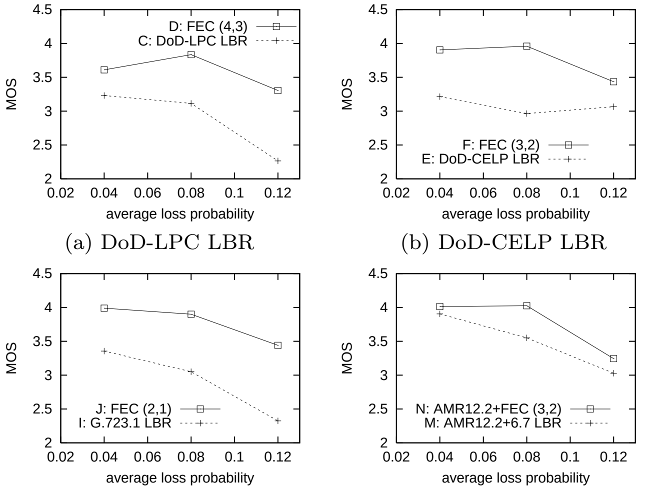
Figure 12: Performance of LBR variants and optimality of our LBR implementation

5.2 Test Set $\mathcal{N}_1$: Quality of FEC vs. LBR

Figure 13 compares the MOS achieved by FEC and LBR over four different codec configurations. The first three use G.729 as the main codec, with DoD-LPC [16], DoD-CELP [3], and G.723.1 [12] as secondary LBR codecs. The last one use AMR [8] 12.2 kb/s as the main codec, and AMR 6.7 kb/s as the secondary LBR codec.

In Figure 13, occasionally the MOS does not decrease as loss probability increases. This is due to the effect of loss location [22], because some parts of voice are more vulnerable to loss than others. It can sometimes cause fewer losses to be sound more harmful. However, this effect does not interfere with our interpretation of the results at all, because the relative trends between FEC and LBR are always very consistent and clear. In all sub-figures of Figure 13, the FEC MOS curves (solid lines) are consistently higher than the LBR MOS curves (dashed lines). The gap is widest for G.723.1 LBR, and narrowest for AMR LBR.

According to Figure 13, FEC has a clear advantage over any LBR configuration. This is probably because FEC recovers lost packet in *bit-exact* form, therefore decoder state drift problem is non-existent unless a packet is unrecoverable. In addition, in LBR each packet loss will cause a sudden switch between high (original) and low quality (redundant) audio. The sudden switching can cause some listening discomfort, which may also be attributed to LBR's lower MOS score.

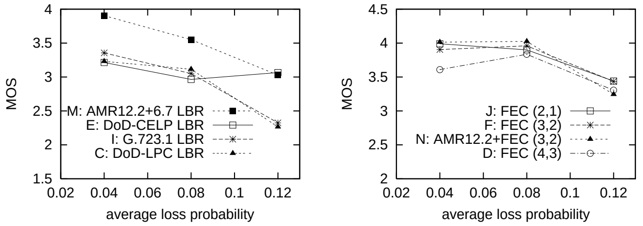


(a) DoD-LPC LBR (b) DoD-CELP LBR

(c) G.723.1 LBR (d) AMR 12.2+6.7 kb/s LBR

Figure 13: FEC vs. LBR on MOS quality, Gilbert loss ($p_c = 30\%$); G.729 is main codec except for case (d); T=30 ms except for case (d) where T=20 ms

Figure 14(a) compares the MOS performance of all four LBR configurations. AMR LBR has the best quality, possibly because both its main and secondary codecs are of the same family (AMR), and this "affinity" may have improved LBR quality. All three other LBRs are similar except DoD-CELP LBR is better at higher loss rate (12%). It should be noted however, that even AMR LBR produces lower MOS than its corresponding FEC code.



(a) LBR based on different codecs (b) different FEC codes

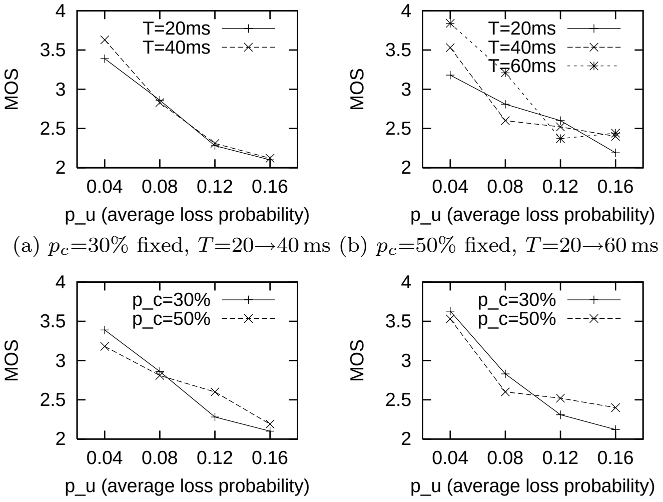
Figure 14: Comparisons within the same category

Among all FEC codes in Figure 14(b), FEC (3,2) (both G.729 and AMR) and (2,1) code has the best quality, (4,3) code is slightly worse, because it has a smaller amount of redundant data. Notice that at 4% loss, the MOS of (4,3) code is lower than MOS at 8% loss. This is not because the FEC code performed badly, in fact, the final loss probability is within expected range. There loss location [22] must have had a special impact on quality, as explained earlier. Because p_f values for both 4% and 8% cases are small and comparable, it is quite possible for the 4% case to sound worse than the 8% case. But again this does not change the fact that FEC produces better quality than LBR.

Finally, in terms of CPU complexity, FEC is much more economical than LBR. For example, G.729 and G.723.1 requires 20 and 14.6 MIPS (million instructions per second)[5], respectively. An FEC code only requires a few logical operations (e.g., XOR) for every payload byte. G.729 runs at 8 kb/s and thus produces 1000 payload bytes per second. Therefore a typical FEC code would need only a few kilo logical operations per second, that is, less than 0.01 MIPS. By comparison, LBR requires the sender to run two codecs simultaneously. Therefore an LBR configuration of G.729 + G.723.1 would require a total of 20+14.6 = 34.6 MIPS, as opposed to 20+0.01 = 20.01 MIPS with FEC. In terms of delay, it depends on the FEC code block size and LBR lag. In our experiments they are generally comparable.

5.3 Test Set $\mathcal{N}_2$: MOS Quality vs. Loss Burstiness and Packet Interval

5.3.1 MOS without FEC



(a) p_c=30% fixed, T=20→40 ms (b) p_c=50% fixed, T=20→60 ms

(c) T=20 ms fixed, p_c=30→50% (d) T=40 ms fixed, p_c=30→50%

Figure 15: MOS quality with respect to loss burstiness (p_c=30-50%, specified at T=20 ms) and packet interval (T=20-60 ms), no FEC case

Figure 15 shows how MOS changes over a loss range from 4% to 16%, under different loss burstiness (p_c) and packet intervals (T). Figure 15(a) and (b) illustrate how packet interval affects MOS. In both (a) and (b), increasing T always improves MOS slightly when loss is low (p_u=4%). A possible explanation is that under bursty loss, with a larger T, there will be fewer loss bursts, hence fewer distortions. When overall loss is low, this may be more appealing to a listener. For higher loss rates, however, the trend is not clear.

Figure 15(c) and (d) show how loss burstiness (p_c) affects MOS. For both (c) and (d), a higher p_c degrades MOS slightly for low loss (p_u=4%), but improves MOS slightly for high loss (p_u=12-16%). To our knowledge, these two opposite trends have not been cited in previous studies [15, 9]. This behavior is also different from the results in section 5.1, but they do not contradict each other, because section 5.1 compares only random and bursty loss (with p_c=30%), whereas Figure 15(c) and (d) compares the MOS effect of different degrees of loss burstiness (with p_c=30% and 50%). Finally, for both T=20 ms and 40 ms, we see the same opposite trends, so this behavior is less likely to be a random artifact of our MOS tests (e.g., selection of random seeds for packet loss). To determine the precise effect loss burstiness has on perceived quality, however, would require considerably more tests.

5.3.2 MOS of FEC vs. Packet Interval T

Most portions of the MOS curves in Figure 15 (where $p_u \geqslant 8\%$) are not very applicable in practice. Users are unlikely to tolerate a service where quality is mostly below fair (i.e., MOS $\leqslant$ 3.0). However, if FEC can greatly improve MOS, this will become a viable service. This has motivated us to conduct more MOS tests with FEC. Since we already know from the results of test set $\mathcal{N}_1$ (section 5.2) that FEC gives much better quality than LBR, it is desirable to find under what setting FEC maximizes MOS quality.

Figure 16 shows how FEC improves quality over Figure 15(a), where loss burstiness p_c is 30%. In particular, the FEC MOS at T=40 ms is consistently 0.5 to 0.6 MOS higher than the corresponding FEC MOS at 20 ms. This is illustrated by the vertical arrow in Figure 16. So the selection of packet interval can be vital on FEC quality. Apparently, a larger packet interval wins in this case.

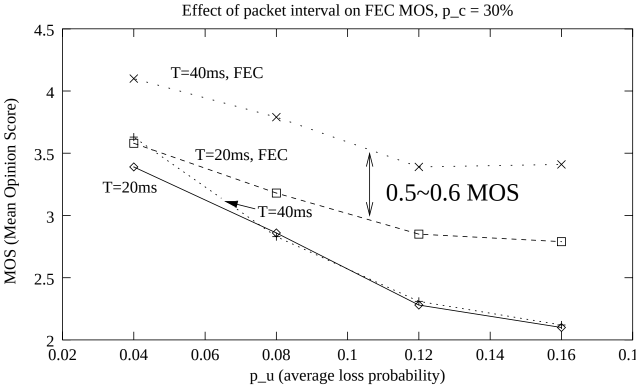
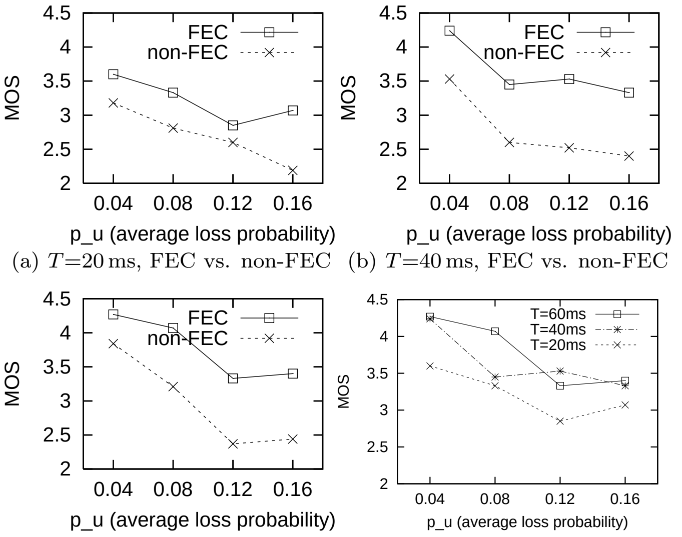


Figure 16: MOS quality of FEC under different packet intervals, (3,2) piggyback FEC code, p_c = 30% specified at T=20 ms

Figure 17 is similar to Figure 16, except that loss is burstier (p_c = 50% instead of 30%). Figure 17(a), (b) and (c) compare the MOS of FEC against non-FEC, and (d) compares FEC among different packet intervals. From Figure 17(d), we see that similar to Figure 16, increasing T also gives better FEC MOS. The only small exception is at p_u=12%, when T changes from 40 to 60 ms, FEC MOS decreases slightly. We have verified that the final loss probability after FEC (p_f) is within expectation for both 40 and 60 ms packet intervals. There should not be any significant measurement error in the MOS result data either, because we can clearly see in both Figure 17(b) and (c) a consistent and near-constant increase in MOS from the non-FEC curve to FEC curve. Such consistency is unlikely if the MOS results are unreliable or have large variance.



(a) T=20 ms, FEC vs. non-FEC (b) T=40 ms, FEC vs. non-FEC

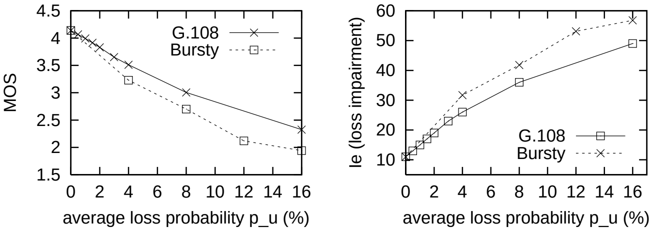
(c) T=60 ms, FEC vs. non-FEC (d) Comparing FEC, T=20-60 ms

Figure 17: MOS quality of FEC under different packet intervals, (3,2) piggyback FEC code, p_c = 50% specified at T=20 ms

A plausible explanation is that with a larger packet interval, the duration of final loss bursts (which is proportional to T) is also much longer. Fewer but much "bigger" losses could result in worse quality, especially if an average burst is long enough to wipe out an entire phoneme in the speech. Another possibility is due to the effect of loss location [22], as explained in section 5.2. To determine which is the actual cause would require many more MOS tests.

Figure 17(d) suggests that for low loss (p_u=4-8%), FEC has highest MOS with a 60 ms packet interval, and for high loss (12-16%), a packet interval of 40 ms may be preferable. This may serve as a guidance on how to select the best packet interval.

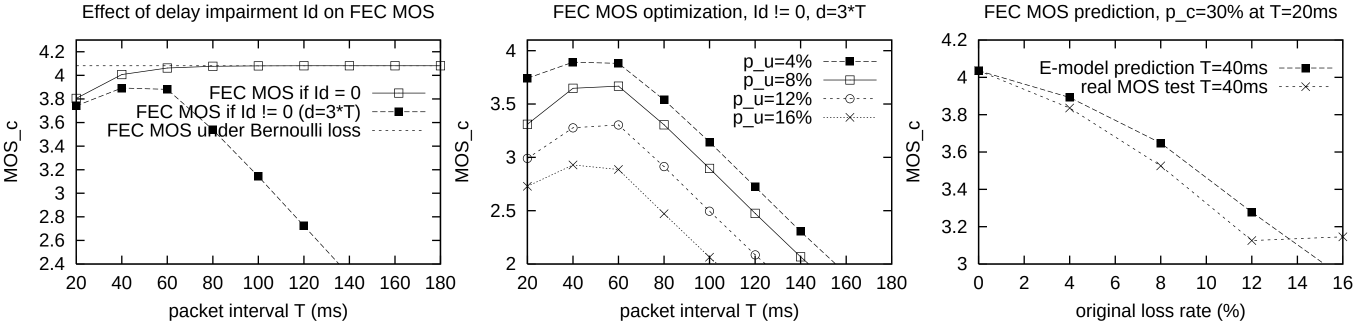
5.3.3 Comparison with the E-model



(a) MOS curves (b) loss to I_e mapping

Figure 18: G.108 mapping (random loss) vs. our MOS results (bursty loss, p_c = 30%), G.729 codec, T=20 ms

For the G.729 codec, the E-model supplies only a random loss mapping in the ITU G.108 recommendation [15], with a packet interval of 20 ms, as in Figure 9. Since one of our MOS tests in set $\mathcal{N}_2$ has the same configuration except it is under bursty loss (p_c=30%, as in Figure 15(a)), we compare both of them in Figure 18(a). We have normalized our MOS results by having the same starting point at 0% loss, so that



(a) MOS_c vs T, p_u=4% (b) MOS_c vs T, p_u=4-16% (c) Validating prediction of E-model

Figure 19: Trade-off between final loss probability (p_f) and delay impairment (I_d) when optimizing packet interval T to maximize FEC *conversational* MOS; p_c=30%; using (3,2) FEC code with piggyback

it is a fair comparison to the E-model mapping. Similar to the two G.711 curves in Figure 9, bursty loss (our result curve) has lower quality than random loss (G.108 curve).

5.3.4 Optimizing Packet Interval with Delay Impairment

Normally a MOS test is listening only and does not consider the effect of delay. The resulting MOS is also called *listening* MOS. In practice, high delay impairs interactive conversations. The E-model maps this effect to the delay impairment score (I_d) and then calculates a *conversational* MOS, denoted as MOS_c. Because the R to MOS mapping in Figure 8(b) is reversible, we can reversely map MOS into the corresponding R value, and then derive the corresponding loss impairment score (I_e) as $94.2 - R - I_d$, per Equation (5). Apparently I_d should be set to 0 if the MOS results were listening-only. Then we can obtain a loss rate to loss impairment score mapping. Figure 18(b) illustrates this reverse-engineered loss (p_u) to impairment (I_e) mapping for G.729 bursty loss (p_c=30%). It is compared against G.108's random loss to I_e mapping. Apparently the impairment score is higher under bursty loss.

Figure 6(c) shows how the final loss probably p_f decreases when the packet interval T increases. If we use Formula (4) to predict p_f from the original loss probability p_u, and then apply the loss to I_e mapping in Figure 18(b), we can use the E-model to predict FEC MOS. If delay impairment is not considered, that is, $I_d = 0$, the FEC MOS predicted by the E-model will always increase with T. However, for an (n, k) FEC code, the maximum FEC delay d introduced is nT including the packetization delay. So the delay impairment I_d is no longer zero. Figure 19(a) demonstrates the effect of delay impairment on predicted FEC *conversational* MOS (MOS_c) for a (3,2) FEC code with piggybacking. Here a packet interval of 40 ms maximizes MOS_c, when the original loss is 4%.

Applying the same procedure, we can obtain the trend of FEC MOS_c for other original loss rates, as illustrated in Figure 19(b). Here FEC MOS_c is maximized with a packet interval of 40 or 60 ms, and the MOS difference between using a 40 ms and 60 ms interval is minor.

To validate MOS predictions of the E-model in Figure 19(b), we plot the predicted FEC MOS at T=40 ms versus its reference FEC MOS curve from Figure 16. The reference curve is normalized so that it also takes delay impairment into account. The resulting comparison chart is Figure 19(c). Overall, the prediction curve matches well with the real test. The MOS predicted by the E-model is slightly higher than our actual test results. This is possibly because the final loss pattern after FEC is much burstier than the original loss, since the unrecoverable packets are apparently those that come in long bursts. Therefore, the p_u to I_e mapping that we use here in Figure 18(b) (which is based on its original and less bursty loss pattern) may no longer apply accurately. This may require re-doing the loss to I_e mapping and it is part of our future work.

The test curve in Figure 19(c) becomes horizontal between 12 and 16% loss. The exact reason for this is unknown, but it is possible users become insensitive to quality drop beyond a certain loss rate. It requires more MOS testing to determine the right cause.

6. CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

We present an evaluation study on the effect of random and bursty (Gilbert) packet loss on VoIP perceived quality. The results indicate that bursty loss leads to lower MOS than random loss, but the trends sometimes become less clear if loss burstiness is too high. Next, we conduct more MOS tests to compare the quality achievable by FEC and low bit-rate redundancy (LBR). Our results show that FEC is prefer to LBR for all LBR codec configurations tested. Finally, both our theoretical analysis and real MOS tests confirm that a larger packet interval generally improves FEC quality. We then describe a new method that maximizes FEC *conversational* MOS by choosing an optimal packet interval. The method is based on the E-model and it automatically trades off between FEC delay and listening quality.

Our MOS tests explore a few areas not covered in previous studies [14, 15]. This includes a wider range of application settings, such as the choice of packet interval and the degree of loss burstiness. Our MOS comparison study of FEC and LBR is the first to demonstrate FEC's advantage over LBR. Although there have been studies on FEC objective performance, for example, in terms of final loss probability, our investigation of FEC MOS, or subjective performance is unique. Finally combining final loss probability prediction and the E-model loss impairment mapping, we have established a procedure for selecting the optimal packet interval that maximizes conversational MOS (MOS_c) under FEC.

6.2 Future Work

Some of our listening test results have certain out-lier points that are not uniquely explainable. An example is Figure 17(d), where the FEC curves for T=40 and 60 ms are not very regular. It could be due to either loss location effect [22], or the fact that the average final loss burst is too long with large Ts. Determining which is the real reason requires more MOS tests. By averaging over many tests, the effect of loss location can be minimized, but the effect due to average loss burst length would not be.

Loss impairment depends a lot on the loss pattern. We have seen in Figure 18 that burstier loss usually leads to lower quality. The final loss pattern after FEC is generally much burstier than its original loss pattern, because those packets that are unrecoverable usually come in long bursts. Therefore, we need more MOS testing to re-calibrate new loss to impairment (I_e) mappings for accurate prediction of FEC MOS.

Finally, since FEC is clearly the preferred loss repair method, we plan to evaluate the trade off among the bandwidth overhead and delay an FEC code introduces, and the amount of MOS improvement FEC brings. The corresponding results will shed light on how much overhead we need to achieve a certain level of QoS (quality of service).

7. REFERENCES

[1] Bellamy, J. C. *Digital Telephony*, third ed. John Wiley & Sons, Inc., 2000.

[2] Bolot, J.-C., and Garcia, A. V. Control mechanisms for packet audio in the Internet. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)* (San Fransisco, California, Mar. 1996).

[3] Campbell, Jr., J. P., Welch, V. C., and Tremain, T. E. The new 4800 bps voice coding standard. In *Military and Government Technology* (Arlington, Virginia, Nov. 1989).

[4] Cole, R. G., and Rosenbluth, J. Voice over IP performance monitoring. *ACM Computer Communication Review 4*, 3 (2001).

[5] Cox, R. V., and Kroon, P. Low bit-rate speech coders for multimedia communication. *IEEE Communications Magazine 34*, 12 (Dec. 1996), –.

[6] Dynastat, I. Tdoc. s4 (00)0587, 3G AMR-NB characterization experiment 1A - Dynastat results. Tech. rep., 3GPP, Dec. 2000.

[7] Frossard, P. FEC performance in multimedia streaming. *IEEE Communications Letters 5*, 3 (Mar. 2001), 122–124.

[8] Group, T.-S. C. W. 3G TS 26.091, AMR speech codec; error concealment of lost frames. Tech. rep., 3GPP, 1999.

[9] group T1A1.7, W. Results of a subjective listening test for G.711 with frame erasure concealment. Tech. rep., Committee T1, May 1999.

[10] Hardman, V., Sasse, A., Handley, M., and Watson, A. Reliable audio for use over the Internet. In *Proc. of INET'95* (Honolulu, Hawaii, June 1995).

[11] International Telecommunication Union. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction. Recommendation G.729, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Mar. 1996.

[12] International Telecommunication Union. Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. Recommendation G.723.1, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Mar. 1996.

[13] International Telecommunication Union. Subjective performance assessment of telephone-band and wideband digital codecs. Recommendation P.830, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Feb. 1996.

[14] International Telecommunication Union. The e-model, a computational model for use in transmission planning. Recommendation G.107, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Dec. 1998.

[15] International Telecommunication Union. Application of the e-model: A planning guide. Recommendation G.108, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Sept. 1999.

[16] National Communications Systems. Analog to digital conversion of voice by 2,400 Bit/Second linear predictive coding (federal standard 1015). Tech. Rep. FED-STD-1015, General Services Administration, Nov. 1984.

[17] Perkins, C., Hodson, O., and Hardman, V. A survey of packet loss recovery techniques for streaming audio. *IEEE Network 12*, 5 (Sept. 1998), 40–48.

[18] Perkins, C., Kouvelas, I., Hodson, O., Hardman, V., Handley, M., Bolot, J. C., Vega-Garcia, A., and Fosse-Parisis, S. RTP payload for redundant audio data. Request for Comments 2198, Internet Engineering Task Force, Sept. 1997.

[19] Rosenberg, J., Qiu, L., and Schulzrinne, H. Integrating packet FEC into adaptive voice playout buffer algorithms on the Internet. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)* (Tel Aviv, Israel, Mar. 2000).

[20] Rosenberg, J., and Schulzrinne, H. An RTP payload format for generic forward error correction. Request for Comments 2733, Internet Engineering Task Force, Dec. 1999.

[21] Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V. RTP: a transport protocol for real-time applications. Request for Comments 1889, Internet Engineering Task Force, Jan. 1996.

[22] Sun, L., Wade, G., Lines, B., and Ifeachor, E. Impact of packet loss location on perceived speech quality. In *Internet Telephony Workshop 2001* (New York, Apr. 2001).

[23] Yajnik, M., Moon, S., Kurose, J., and Towsley, D. Measurement and modelling of the temporal dependence in packet loss. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)* (New York, Mar. 1999).