Perceived Quality of Packet Audio under Bursty Losses

Wenyu Jiang, Henning Schulzrinne
Columbia University
{wenyu,hgs}@cs.columbia.edu

Abstract—**We examine the impact of bursty losses on the perceived quality of packet audio, and investigate the effectiveness of various approaches to improve the quality. Because the degree of burstiness depends on the packet interval, we first derive a formula to re-compute the conditional loss probability of a Gilbert loss model when the packet interval changes. We find that FEC works better at a larger packet interval under bursty losses. In our MOS-based (Mean Opinion Score) listening tests, we did not find a consistent trend in MOS when burstiness increases if FEC is not used. That is, in some occasions MOS can be higher with a higher burstiness. With FEC, our results confirms the analytical results that quality is better with a larger packet interval, but T should not be too large to avoid severe penalty on a single packet loss. We also find that low bit-rate redundancy generally produces lower perceived quality than FEC, if the main codec is already a low bit-rate codec. Finally, we compare our MOS results with objective quality estimation algorithms (PESQ, PSQM/PSQM+, MNB and EMBSD). We find PESQ has the best linear correlation with MOS, but the correlation is still not high enough to be used in isolation to predict MOS.**

Keywords—**packet audio, IP telephony, voice over IP, perceived quality, subjective quality, loss model, Gilbert model, forward error correction, low bit-rate redundancy.**

I. Introduction

IP telephony based on packet audio has drawn significant interests due to the potential cost savings and new services [27] it can offer. When deploying it in the current Internet, however, one must consider the effect of packet loss and delay. Packet losses in the Internet are temporally correlated [4], [3], [19], [26], that is, they often come in bursts rather than with a random (i.e., Bernoulli) pattern. It is therefore useful to study how this affects the perceived quality of packet audio and its quality improvement measures such as Forward Error Correction (FEC) and low bit-rate redundancy [20]. In our listening tests, we use both Internet packet traces and the Gilbert model [17].

The Gilbert model has two parameters, unconditional and conditional loss probability, which we denote as p_u and p_c, respectively. p_c is defined as the probability that the next packet will be lost given the previous one was lost. p_c is a simple yet useful measure of burstiness, but for a given network path, its value clearly depends on the packet interval T. Since loss is generally due to router buffer overflow, sending packets more frequently during that time will lead to more *consecutively* lost packets, hence higher p_c.

However, the Gilbert model is often used without reference to the packet interval T. Consequently, results from one experiment may not be applicable to another if T is different. We address this issue by presenting an exact formula that recalculates p_c upon change of T.

Then we analyze the final loss rate of FEC under a Gilbert model. The results indicate that FEC works better at a larger T, but apparently T cannot be arbitrarily large, as its adds more end-to-end delay. Also, as T increases, the penalty of a single packet loss can become intolerable. According to Hardman [9], a phoneme in human speech is on average 80 ms. Therefore, any T near or beyond 80 ms will cause a single packet loss to eliminate the whole phoneme, thus reducing intelligibility significantly.

To verify the analytical results and how large T should be, we perform a series of subjective listening tests. The results show that, the quality with FEC is generally better at a larger T, but for some conditions the MOS is not necessarily higher at 60 ms than 40 ms for higher loss rates. We also find that loss recovery using low bit-rate redundancy generally has a lower perceived quality than FEC.

Finally, we compare our MOS results with several objective quality estimation algorithms (PESQ, PSQM/PSQM+, MNB, and EMBSD). In most occasions, these algorithms agrees with subjective MOS in term of "better or worse" judgment, but the actual prediction of MOS is less effective. Among them, PESQ has the highest linear correlation with subjective MOS, but still they are not precise enough to directly predict MOS. Furthermore, MNB has a strong saturation effect in some cases, that is, MNB almost always thinks the audio clips are very good.

The rest of the paper is organized as follows. Section II derives the formula for recalibrating Gilbert parameters. Section III analyzes FEC performance under the Gilbert model. Section IV presents the design methodology of our listening tests. Section V discusses the test results. Section VI studies the MOS correlation of the objective quality measurement algorithms. Section VII lists related work in subjective quality evaluation under packet losses. Section VIII summarizes the results.

II. Quantifying the Effect of Down-sampling and Up-sampling in the Gilbert Model

The quality of packet audio depends strongly on the network loss pattern. An analytical model such as the Gilbert Model is often used to approximate a packet loss trace in perceived quality evaluations. However, when the packet interval changes, the conditional loss probability (p_c) of a Gilbert model may also change. For example, if a packet trace is created with a 30 ms packet interval, and has a p_u (overall loss) of 5% and p_c of 30%, it would be incorrect to use the same p_c to simulate the path behavior at 10 ms or 60 ms packet intervals. This requires a recalibration of p_c, but for packet traces that already exist, it is impossible to collect the trace again. Even for traces that are yet to be created, having to use the "lowest common denominator" (e.g., 10 ms) packet interval results in very large trace files. Therefore, an analytical recalibration of p_c is preferred.

When the packet interval T increases or decreases, it can be thought of as a form of down-sampling and up-sampling. Next we will study the dependence Gilbert model parameters on T from this perspective.

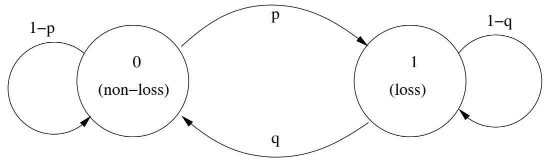


Fig. 1. The Gilbert model specified with p and q

A Gilbert model can be specified in two ways: p_u and p_c, or with state transition probability p and q, as illustrated in Figure 1. State 0 is non-lossy, whereas state 1 is lossy. A transition from state 0 to 1 is a packet loss, so is a 1-1 transition. Apparently, $p_c = 1 - q$. If we denote the state resident probability as π_0 and π_1, then π_1 is simply p_u. As in [17], their relationships are:

$$\pi_0 = \frac{q}{p+q}, \quad \pi_1 = \frac{p}{p+q} = p_u$$

which can be transformed into:

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

The loss burst length k in a Gilbert model has a geometric distribution: $p_k = (1-q)^{k-1} \cdot q$. We can then derive the mean loss burst length $E[k]$ as:

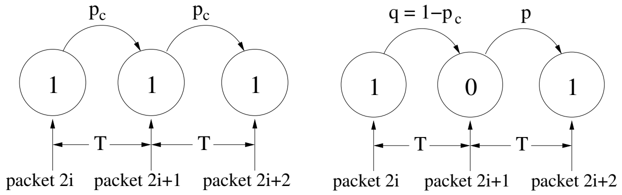
$$E[k] = \sum_{k=1}^{\infty} p_k \cdot k = \frac{1}{q} = \frac{1}{1 - p_c} \qquad (2)$$

At a first glance, re-calibration of p_c is simple. The Gilbert model has a counter-part in the continuous domain: the continuous 2-state Markov chain. The system is in a non-loss (0) or lossy (1) state for an exponentially distributed period of time. Its parameters are the average durations in each state, denoted as τ_0 and τ_1. They do not depend on the sampling period. All one has to do is measure τ_0 and τ_1, and map them to the discrete Gilbert model.

However, it is difficult to measure or infer precisely the parameters of a continuous model using discrete measurements. If we estimate τ_1 to be the average duration of a loss burst, i.e., $E[k] \cdot T = 1/(1 - p_c) \cdot T$, we will find that τ_1 cannot be constant irrespective of T. For example, if at T = 20 ms, p_c = 30%, then $\tau_1 = 1/0.7 \cdot 20$ ms = 28.6 ms. Then at T = 40 ms, even if p_c = 0%, τ_1 can only be as low as 40 ms. Therefore we cannot use discrete measurements ($E[k] \cdot T$) to precisely estimate τ_1.

A. T to $2T$ Down-Sampling

We first ask: what happens to p_c when T is changed to $2T$? We denote the new p_c as either p'_c or $p_{c,k}$, where $k = 2, 3, 4, ...$ is the T to kT down-sampling ratio.



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

Fig. 2. Down-sampling of a Gilbert process ($T \to 2T$)

During down-sampling, in $2T$ time scale, if packet i is already lost, and packet $i+1$ is also lost, it counts as a sample of p'_c. If we look at the original loss sequence in T time scale (Figure 2), it corresponds to the loss of packet $2i$ and $2i+2$. The question now is whether packet $2i+1$ have been lost. The probability of packet $2i+1$ also lost is p_c (Figure 2a), and chance of it not lost is q (Figure 2b). So p'_c is now calculated as:

$$p'_c = {p_c}^2 + q \cdot p = {p_c}^2 + (1 - p_c) \cdot p$$

Using Formula (1) to substitute p, we get

$$p'_c = {p_c}^2 + \frac{(1 - p_c)^2 \cdot p_u}{1 - p_u} \qquad (3)$$

In $2T$ time scale, if the system is in state x ($x \in$ {0,1}), its transition probability only depends on its current state x, and does not depend on its previous states. This is because of the memoryless property of the Gilbert process at T time scale. Therefore, the new loss process is still memoryless at $2T$, hence, still a Gilbert process.

B. General T to kT Down-Sampling

For $k = 3$, there are totally 4 cases to consider, as shown in Figure 3. Its exact formula after summing them up is:

$$p_c' = {p_c}^3 + q(1-p)p + q \cdot p \cdot p_c + p_c \cdot q \cdot p$$

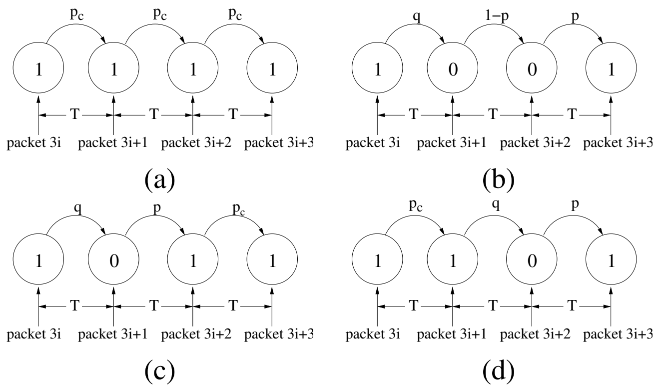


Fig. 3. Enumeration of cases for k=3

After simplification and factoring, it becomes

$$p_c' = \frac{{p_c}^3 + p_u - 2{p_u}^2 + 3p_u \cdot p_c - 3 \cdot p_u \cdot {p_c}^2}{(1-p_u)^2}$$

$$= \frac{(p_c - p_u)^3 + p_u \cdot (1-p_u)^2}{(1-p_u)^2} = \frac{(p_c - p_u)^3}{(1-p_u)^2} + p_u$$

This expression does not look identical to Formula (3), but we can transform Formula (3) into

$$p_c' = {p_c}^2 + \frac{(1-p_c)^2 \cdot p_u}{1-p_u} = \frac{(p_c - p_u)^2}{1-p_u} + p_u$$

Therefore we conjecture the generalized formula to be:

$$p_{c,k} = \frac{(p_c - p_u)^k}{(1-p_u)^{k-1}} + p_u \quad (4)$$

where $p_{c,k}$ is the conditional loss probability at kT sampling. We can easily prove that the new process under kT sampling period is still a Gilbert process. This is because at any kT time scale, the system's transition only depends on its current state, not any earlier states, due to the memoryless property at T time scale. To prove Formula (4), we use induction.

[Proof] The formula already holds for $k = 2, 3$. If the statement is true for $k-1$, we just need to prove,

$$\frac{p_{c,k} - p_u}{p_{c,k-1} - p_u} = \frac{p_c - p_u}{1 - p_u}$$

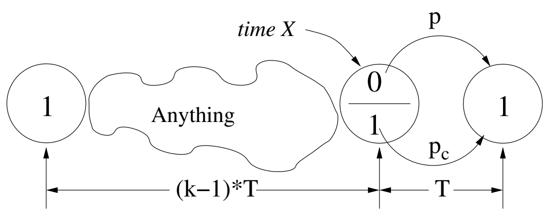
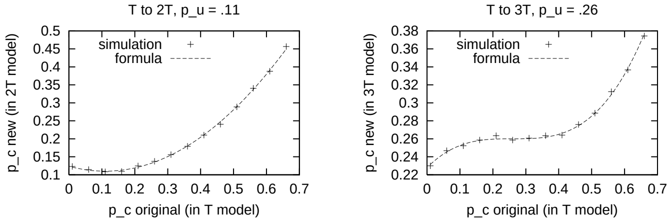


Fig. 4. Proof by induction for general $T \rightarrow kT$ case

From Figure 4, the probability that at time X the system state is 0 is: $1 - p_{c,k-1}$, therefore



(a) $T \rightarrow 2T$, p_u=11% (b) $T \rightarrow 3T$, p_u=26%

Fig. 5. Simulation of sub-sampling in the Gilbert model

$$p_{c,k} = (1 - p_{c,k-1}) \cdot p + p_{c,k-1} \cdot (1-q)$$
$$= p + p_{c,k-1}\left(\frac{-p_u(1-p_c)}{1-p_u} + p_c\right)$$
$$= p + p_{c,k-1}\frac{(p_c - p_u)}{1-p_u}$$

since $p = \frac{p_u \cdot (1-p_c)}{1-p_u} = \frac{p_u(p_u - p_u + 1 - p_c)}{1-p_u}$

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

Q.E.D.

Formula (4) predicts $p_{c,k}$ as a power function of p_c, with a minima at $p_c = p_u$ if k is even. To verify our formula, we run a down-sampling simulation on the Gilbert model, with p_c as the independent variable. The results are shown in Figure 5. The simulation clearly confirms the formula's correctness. Also notice the part of the curves where $p_c < p_u$, that is, loss is less bursty than Bernoulli. They have different trends depending on whether k is even or odd. In practice, such "un-bursty" loss patterns do not often occur.

C. Arbitrary Up-Sampling and Down-Sampling

Formula (4) calculates the new p_c in the case of down-sampling, i.e., a packet interval increase in integer ratios, but it can be extended to up-sampling as well. There is no requirement that if a process is Gilbert at kT, it has to be Gilbert as well at T. However, if we assume the loss process is close enough to be Gilbert at T, we can reverse Formula (4) as follows:

$$(p_c - p_u)^k = (p_{c,k} - p_u) \cdot (1-p_u)^{k-1}$$

If k is even, there are two real roots, one positive and the other negative. But if we can safely assume that burstiness only goes up during up-sampling, which is a reasonable assumption, only the positive root will be applicable. Then:

$$p_c = p_u + \sqrt[k]{(p_{c,k} - p_u) \cdot (1-p_u)^{k-1}} \quad (5)$$

This is useful, for example, when we approximate a 30 ms based packet trace with a Gilbert model, but we want to know the path behavior if it were transmitting at 10 ms. In fact, if we can up-sample it to 10 ms without introducing much error, we can also perform a down-sampling from 10 ms to 20 ms. So using a concatenation of Formula (4) and (5), we can estimate the network behavior at nearly arbitrary time scale, assuming the network can be safely modeled as a Gilbert process.

In fact, it is not difficult to generalize that in any $T \to kT$, where k is any positive real number, Formula (4) holds. In addition, Formula (5) is simply another way of writing Formula (4), where k becomes $1/k$.

III. Analytical Performance of FEC under the Gilbert Loss Model

An FEC code improves transmission quality by sending redundant data in addition to the payload. Its most common types are parity and Reed-Solomon (RS) code [2]. A parity or RS code is referred to as a (n, k) code if its block size is n units and payload is k units. It works by sending data in blocks, first k units of payload, then $n - k$ units of redundant data. If no less than k units in a block is correctly received, all the payload in this block can be recovered. In the case of packet audio, each unit is an audio packet, sent regularly at the packet interval.

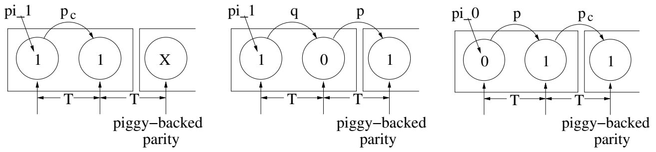
Given the same p_u and FEC code, it is evident that if losses occur in long bursts (length $> k$), the FEC block will not be able to recover them. Therefore FEC is less effective under bursty losses compared to random (i.e., Bernoulli) losses.

The final packet loss probability, p_f, of an (n, k) FEC code under Bernoulli losses [22], where $p = p_u$, is:

$$p_f = p \left(1 - \sum_{i=k}^{n-1} \binom{n-1}{i} (1-p)^i p^{n-1-i} \right) \qquad (6)$$

FEC's performance under bursty losses is more complex. Frossard [6] gives an in-depth derivation of FEC performance parameters in any renewal error process, including a Gilbert loss process. The performance parameters are the final packet loss probability p_f and average final loss burst length. The results of [6] assume the typical transmission scheme in packet video or telecommunications network, where the FEC blocks are sent sequentially without any overlapping. Packet audio, however, usually use a "piggy-back" scheme, where the FEC packets are sent on the first few packets of the next FEC block. This reduces the total number of packets and the protocol header overhead. In our listening tests the FEC clips are generated using piggy-back.

The piggy-back scheme alters the resulting loss process, so we cannot directly apply results of [6] to our listening tests. In addition, the formulas in [6] are not in a simple, closed form, but rather as a set of recurrences and iterations. So we will present here a formula for a simple (3,1) piggy-backed parity FEC code. It has a 50% overhead, and it is one of the FEC codes used in our listening test sets.



(a) case 1 (b) case 2 (c) case 3

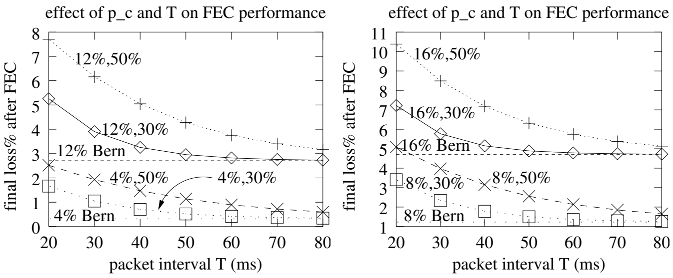
Fig. 6. 1/2 parity FEC under a Gilbert loss process

Given any FEC block, the probability that its first packet is lost is π_1, i.e., the Gilbert state probability or p_u. Using Figure 6(a), the proportion of final lost packets in case 1 is $\pi_1 \cdot p_c \cdot 100\%$. X means "don't care." The 100% indicates both payload packets are unrecoverable.

Similarly, the value for case 2 is $\pi_1 \cdot q \cdot p \cdot 50\%$, the 50% indicates half of the payload is unrecoverable. The proportion for case 3 is $\pi_0 \cdot p \cdot p_c \cdot 50\%$, where π_0 is simply $1 - \pi_1 = 1 - p_u$. The sum of these values is then the final packet loss probability after applying FEC, p_f:

$$p_f = \pi_1 \cdot (p_c + q \cdot p/2) + \pi_0 \cdot p \cdot p_c/2 \qquad (7)$$

Figure 7 shows p_f after applying the (3,1) piggy-backed parity FEC. p_c is specified at T=20 ms. The horizontal lines represent p_f in a Bernoulli model (abbreviated as "Bern" in the figure) Apparently, as T increases, p_f in a bursty Gilbert model approaches the Bernoulli limit.



(a) p_u = 4% and 12% (b) p_u = 8% and 16%

Fig. 7. Final loss rate with a (3,1) parity FEC

Figure 7 and Formula (7) suggests that a larger T is always better for FEC when dealing with bursty losses. There are, however, other considerations that calls for a smaller T. First, T is also the packetization delay, which adds to end-to-end delay. Second, the application must use an adaptive FEC playout algorithm [22] to avoid having to always wait on an entire FEC block. But during a lossy period, even the adaptive FEC playout algorithm has to wait for at least one FEC block, which is even larger than T. Finally, with a large T, a single packet loss can wipe out an entire phoneme [9], reducing speech intelligibility significantly. In Section V, we will discuss our listening test results and among other objectives, verify whether a larger packet interval gives better perceived quality, and if so, how large it should be.

IV. Subjective Listening Test Methodology

A. Choice of Codec, FEC, and Low Bit-rate Redundancy

For our tests, we need to choose a voice codec, some FEC code, and also a low bit-rate redundancy scheme. Then audio clips are encoded with the selected codec and network conditions for listener's evaluation.

G.711 [15] is a low complexity codec widely used in PSTN, providing toll quality at 64 kb/s. Although a perfect choice for an intranet where bandwidth is usually abundant, the relatively high bit-rate makes it less suitable for WAN telephony. We have picked G.729 [11] in our test because it is an ITU standardized low bit-rate codec providing near toll quality. Its characteristics have been extensively studied and are representative of many CELP [2] type codecs. G.729 runs at 8 kb/s with a 10 ms frame.

The amount of FEC redundancy and its block size determines its robustness to losses. The block size also determines the delay introduced by FEC blocks. For the purpose of bandwidth conservation, FEC redundancy should not be too high. To reduce delay, FEC blocks should be small. In our experiment, we use a simple (n,1) parity FEC code, with a block size of n and redundancy overhead of $1/(n-1)$. We use piggy-back in our FEC code. We evaluated both (3,1) and (4,1) cases.

A different form of "FEC" used in [4] is actually low bit-rate redundancy [20]. It works by transmitting a lower bit-rate version of the same audio signal, but piggy-backed to later packets. If a packet with the main audio codec is lost, the lost audio information is substituted with a lower quality version. It can be viewed as a form of loss concealment [21], which replaces the lost waveform with an approximation. In both [4] and [20], the illustrated main codec is a high bit-rate one like G.711 or ADPCM/DVI. Since we choose to use the 8 kb/s G.729 as the main codec, this has several implications:

First, the redundant stream should not be coded at a higher bit-rate. This leaves few choices among standard codecs, including: DoD LPC-10 (2.4 kb/s), DoD CELP (4.8 kb/s), G.723.1 (5.3 or 6.3 kb/s). G.723.1 [12] is not very suitable due to its high CPU complexity, DoD CELP's bit-rate is also relatively high. Therefore we choose LPC-10 as the redundant codec. This corresponds to a 2.4/8 = 30% overhead, comparable to the (4,1) parity FEC code that we will also evaluate.

Second, because each frame in most frame-based codecs including G.729 is not coded independently of previous frames, they suffer decoder state drift during a frame loss. We have designed our simulation program to use LPC-10's decoded waveform to "repair" G.729's decoder state during a packet loss. To do this, at the receiver, we re-encode the LPC-10 waveform into a G.729 coder to produce a "pseudo" G.729 frame and help restore the decoder state. But it is difficult to quantify how well the "repair" is. Moreover, LPC itself is also subject to loss impairment.

Third, using a lower quality redundant stream causes perceived quality to suddenly drop during packet losses, which may introduce an unnatural distortion.

In brief, we have selected the following:

codec	FEC code	Low bit-rate redundancy
G.729	(4,1) & (3,1)	DoD LPC-10 (2.4 kb/s)

B. Experiment Design and Test Procedure

Next, we describe how the audio clips are obtained, processed and graded.

All the original audio clips are either taken from the TIMIT [7] speech recognition database, or recorded by native English speakers using Harvard sentences [24]. In either case, the sentences are phonetically balanced. The clips are adjusted to an active speech level of -26 dBov, as required in ITU standard P.56 [10]. Most clips are 7-10 seconds long, consisting of 3 short, unrelated sentences. This conforms to P.830 [13]. Longer clips would have more packets and be more statistically reliable, but they also increase the listening and judging effort, and very often listeners cannot remember the details of a long clip, due to the human memory recency effect [5], [1].

The degraded clips are created by running the G.729 codec on the original clips, using both a packet trace and a Gilbert model. The tests cover network conditions with various degrees of lossiness (p_u) and burstiness (p_c). The loss model is identified either as a trace number identifying a particular packet trace, or as a p_u, p_c pair. For example, 10%,36% means p_u=10%, p_c=36%. When generating packet losses, we calibrated the random seeds carefully to make the achieved loss rate close to the target value p_u.

When a packet loss occurs, one of three loss repair mechanisms is used to create the degraded clips. The first is "plain," G.729's default loss concealment algorithm. The second one is FEC, and packets unrecoverable by FEC are also masked by the G.729 concealment. The third mechanism is LPC, which uses a DoD LPC-10 redundant stream to repair the G.729 decoder state. G.729's default concealment is never invoked in the LPC case.

The listener always uses a headphone to listen to these clips, and is asked to adjust the listening level properly. Using a speaker phone would make it difficult for listeners to sense any distortion in the degraded clips. Listeners hear both original and degraded clips, and then give an opinion score from 1.0 to 5.0, where 5.0 is excellent, 4.0 good, 3.0 fair, 2.0 poor, and 1.0 bad. The averaged value for a particular clip is then the Mean Opinion Score

(MOS). Our MOS scale is the same as the ACR (Absolute Category) test as defined in the ITU standard P.830 [13], but with a small difference: the listeners are asked to give the score at a granularity of 0.1, e.g., 3.7 instead of 3 or 4. Its main purpose is to achieve maximum precision and minimum variance of a MOS value. We have 20 listeners in our experiments. Most of them are graduate students or computer professionals, but none of them have done MOS evaluation before, except the authors.

The table below summarizes how test clips are created.

source	TIMIT, Harvard sentences
speech level	-26 dBov
clip length	7-10 sec
loss model	trace name, or p_u, p_c pair
packet interval T	10-60 ms
grading	MOS scale, 0.1 granularity
loss repair	plain, FEC, or LPC

We designed two sets of listening tests. The first is aimed as a comprehensive test, covering a wide range of conditions (loss models, loss rate) with a limited number of test clips. The second test set measures the effect of burstiness (p_c) and packet interval T on both FEC and non-FEC packet audio.

V. Subjective Listening Test Results

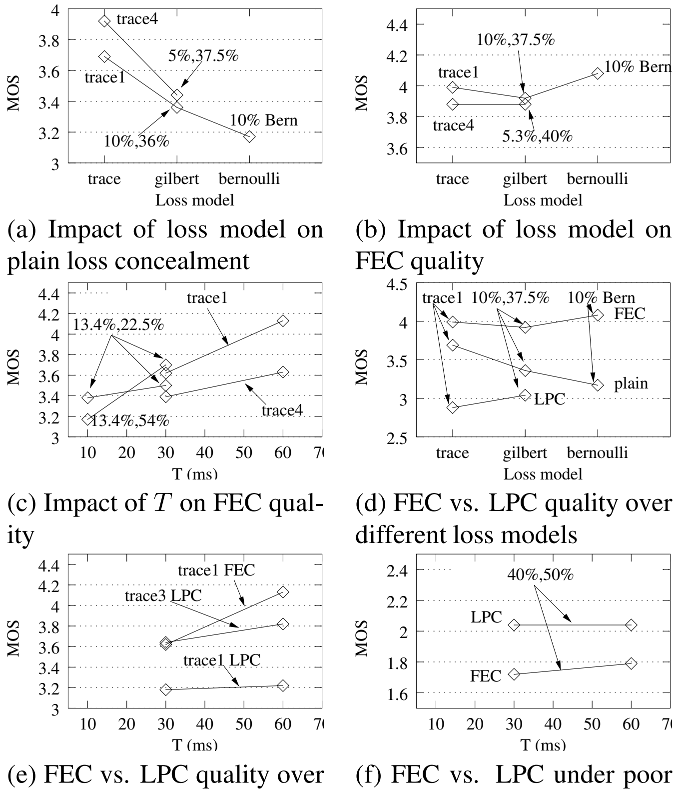
Overall, the test results confirm that using a larger packet interval generally improves the perceived quality with FEC. The results also suggests that FEC has better perceived quality than low bit-rate redundancy.

A. *Results of Test Set 1*

First, the following is a list of packet traces used in the generation of test audio clips. Due to their nature, we can only list their equivalent p_u and p_c to illustrate its burstiness. They are collected between Columbia University, UC Santa Cruz, University of Massachusetts, and GMD Fokus in Germany. The dates of the traces are a bit old, mainly because newer packet traces we obtained have too few packet losses, mostly due to connectivity to Internet-2.

trace	sender	receiver	date	p_u	p_c
1	CU	UMass	9/19/1997	10%	37.5%
2	GMD	UCSC	9/22/1997	13.4%	22.5%
3	UCSC	CU	9/22/1997	5.8%	10.6%
4	UCSC	UMass	9/23/1997	5.3%	39%

Figure 8 provides a visual representation of the results from test set 1. Each data point represents a single test condition, that is, a unique combination of loss model, packet interval T and repair mechanism, as mentioned in the previous section. The notation "10%,36%" means that test condition has a p_u of 10% and p_c of 36%, whereas "10% Bern" means it is a Bernoulli model with p_u = 10%. If



(a) Impact of loss model on plain loss concealment

(b) Impact of loss model on FEC quality

(c) Impact of T on FEC quality

(d) FEC vs. LPC quality over different loss models

(e) FEC vs. LPC quality over different Ts

(f) FEC vs. LPC under poor network conditions

Fig. 8. Impact of loss model and packet interval T on audio quality, Test Set 1

the model is a trace, the trace number is written near that data point. Each line connected by two or more data points share the same p_u. This allows a fair comparison on how loss model, packet interval and repair mechanism affect perceived quality.

Figure 8(a) shows the effect of loss model on G.729's default loss concealment quality. Within either line (i.e., same p_u), the trace models have the highest MOS. Somewhat surprisingly, Bernoulli model has the worst MOS. It is generally thought that burstier losses result in worse quality, and has been validated in [8]. But in this case it is the reverse. We offer the following conjecture: it is likely at about 10% loss rate, a Bernoulli process introduces more short loss bursts, whereas a burstier pattern has fewer, although longer, loss bursts. Since each loss burst maps roughly to one occurrence of distortion, a listener would probably prefer a less frequent occurrence. A longer burst would cause the decoder state to drift more, but since T is 30 ms in clips of group 1, each packet loss would correspond to a 3-frame loss. Rosenberg [23] has shown that G.729 can conceal a single frame loss well, but double and triple frame losses degrades its effectiveness significantly. Therefore, the decoder state drift is probably similar between a 3-frame (1 packet) loss or 6-frame (2 packet) loss. The first authors has informally simulated the same conditions on a longer clip (30 seconds), and his own opinion score also agrees that a 10% Bernoulli pattern has worse quality than a 10%,36% Gilbert pattern. This result does not necessarily contradict earlier results, since many existing studies such as those by ITU often use a T that is equal to the frame size (in this case 10 ms), whereas we use 30 ms here. We have found that the Cisco 7960 IP phones uses a 20 ms packet interval.

Figure 8(b) shows how FEC performance differs between trace, Gilbert and Bernoulli losses. Although the same Bernoulli condition in (a) has lowest MOS, here it has the highest MOS. This is well within expectation, since FEC can recover more packets under non-bursty losses. The MOS value between Gilbert and trace models are very close in both lines, so their quality should be similar.

Figure 8(c) shows how FEC quality changes with the packet interval T. In all four lines, MOS value increases monotonically with T. This strongly confirms the analytical results that under bursty losses, FEC works better with a larger packet interval. The left two lines in Figure 8(c) illustrates what MOS value becomes if the application uses a 10 ms packet interval. The middle left line with a smaller slope makes a "friendly" assumption that at 10 ms interval, p_c is still 22.5% (though in fact it would be higher). Even with this assumption, MOS at T=30 ms is actually slightly higher than 10 ms (3.50 vs. 3.38). So using a smaller packet interval does *not* necessarily increase quality even under this "friendly" scenario. On the other hand, the left line with a higher slope recalibrates p_c for 10 ms using Formula 5, which gives a p_c of 54%, much higher than 22.5%. The MOS difference is also much higher under this more realistic setting (3.77 vs. 3.18). The protocol header overhead is also smaller at T=30 ms than 10 ms. Therefore It is both bandwidth-efficient and FEC-efficient to use a relatively large packet interval.

Figure 8(d) compares the quality of FEC and the LPC repair mechanism. It is evident that LPC (the line at bottom) has a much lower MOS compared to FEC (the line at top), with regard to any loss model. The difference is about 1.0 MOS scale, that is, a difference between "Good" and "Fair" in quality. In fact, its MOS values are even lower than the plain loss concealment. According to some of the listener feedbacks, the LPC clips have unnatural distortions, which makes the listening experience less comfortable and thus gets a lower MOS. The author has looked into this issue, and also compared with the G.711/LPC repair scheme. It seems that with a small packet interval ($<$ 40 ms), similar distortions exist even for G.711/LPC clips. This may be due to the lower MOS quality of LPC-10, which is about 2.4 to 2.5. When a lost packet is repaired (substituted) by a lower quality signal, and each loss burst could cause an unnatural transition in the frequency (perceptual) domain. This problem seems to be worse in the G.729/LPC case than G.711/LPC.

Figure 8(e) compares again FEC and LPC, but at different packet intervals instead of loss models as in (d). Again FEC quality for trace 1 is noticeably better than LPC quality for the same trace. The MOS of FEC increases significantly as T becomes larger, the MOS of LPC also increases with T, but only slightly.

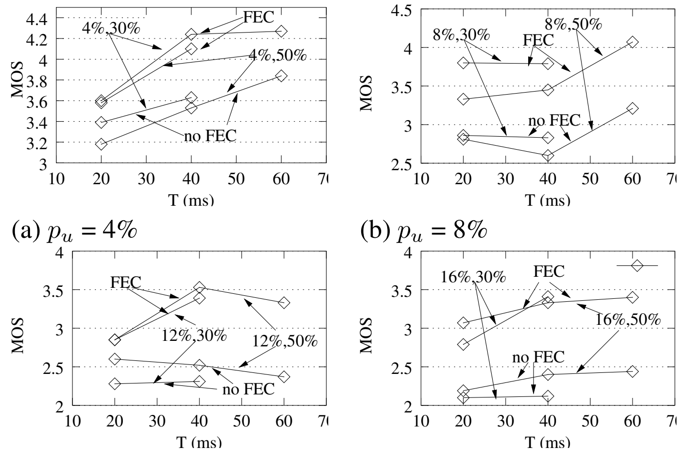
Finally, Figure 8(f) shows performance of FEC and LPC during extremely poor network conditions. Both have 40% loss (with p_c set at a somewhat arbitrary 50%). These are the exceptions where LPC repair work slightly better than FEC. A (4,1) FEC simply cannot repair most lost packets. This results in a final loss rate around 32-33%, and a MOS below 1.8 for both T=30 ms and 60 ms. LPC gives a slightly higher MOS (2.04) in both groups, but the MOS is already on the grade level of "poor." Speech intelligibility is also difficult at this loss rate. The MOS does not change much between 30 ms and 60 ms, probably because of the high loss rate, effect of T is negligible.

To summarize, we find that trace models often produce slightly higher MOS than Gilbert models when only plain loss concealment is used (Figure 8(a)), but they give similar MOS results on FEC (Figure 8(b)). The results confirm that for T = 10, 30, and 60 ms, FEC consistently works at higher Ts, either in final packet loss probability or MOS. LPC repair has worse MOS quality than an equivalent (4,1) FEC code. The only exceptions are when p_u is very high (40%), but the speech in such conditions becomes unintelligible anyway with either LPC or FEC. The worse quality may be due to the lower MOS (2.4-2.5) of the LPC-10 codec. However, the first author has informally tried a low bit-rate redundancy using DoD CELP (4.8 kb/s), which has a much higher MOS of 3.2. The resulting audio has unnatural distortions that are still very audible. It is possible that mixing two different codecs may result in worse quality than each codec would be capable of alone, because many codecs do not preserve the phase information during compression, and mixing two codecs may creates frequent phase changes that become annoying to the ear.

B. *Results of Test Set 2*

Our second test set focus on how different degrees of burstiness affect perceived quality, with or without FEC. p_u increases up to 16%, at a 4% spacing. For each unique p_u, a p_c of 30% and 50% are evaluated, both with and without FEC. For p_c=30%, only two packet intervals are evaluated (20 and 40 ms), but for p_c=50%, three intervals are evaluated (20,40,60 ms). In test set 2, the notation 4%,30% means at T=20 ms, p_u=4% and p_c=30%. When T=40 ms,

p_u remains the same, but the actual value of p_c will be smaller, and can be calculated with Formula (4). We use a (3,1) parity FEC code, giving a 50% redundancy.



(a) $p_u = 4\%$ (b) $p_u = 8\%$

(c) $p_u = 12\%$ (b) $p_u = 16\%$

Fig. 9. Impact of loss burstiness and packet interval T on audio quality, Test Set 2

Figure 9 provides a visual representation of MOS results of test set 2. Notice that all p_c = 30% conditions have only two points (T=20 and 40 ms), whereas all p_c = 50% conditions have three points. Within each sub-figure, p_u is fixed, to allow a simple and fair comparison how burstiness and packet interval affect quality. Each sub-figure contains results for p_c=30% and 50%, with and without FEC. It is easy to see that all the FEC data points and lines are at the top of a sub-figure, and non-FEC ones at the bottom.

When not using FEC, the MOS increases with T when p_u is very low (4%), as in Figure 9(a), but there is no consistent trend when p_u is higher.

In almost all cases, MOS values of FEC at T=20 ms is much lower than at T=40 ms and 60 ms. The only exception is at p_u=8%,p_c=30%, where the 20 ms MOS is nearly the same as 40 ms. The first author looked into this exception, and found that in the T=40 ms clip, most loss bursts hit either on an unvoiced frame or at end of a voiced frame, and some hit the silence period of the audio clip. It has been noted that loss of unvoiced frames or middle/end of voiced frames does less damage to perceptual quality than loss at unvoiced/voiced transitions [28], [25].

Somewhat surprisingly, for p_c=16%,p_c=50%, as in Figure 9(d), the MOS at T=60 ms is not much higher than at 40 ms. In Figure 9(c), where p_c=12%,p_c=50%, the MOS at T=60 ms is even slightly lower than the MOS at 40 ms. In both cases, the final loss rate of the audio clip is close to its theoretical average p_f, so it is not an issue of random seed selection. This is because with a larger T a single loss penalty too severe. Since the average duration of a phoneme in human speech is about 80 ms [9], a single lost packet at T=60 ms can strongly distort a phoneme. A double loss would certainly wipe out the whole phoneme, rendering the speech unintelligible. This exception does not occur in test set 1, mainly because the packet interval in set 1 is compared between 30 ms and 60 ms, rather than 40 ms and 60 ms. Therefore, it is not always better to use a larger T, one has to consider the severity of a single loss penalty as well. According to the results in test set 2, the rule of thumb seems to favor a 40 ms packet interval at higher loss, and prefer 60 ms at lower loss.

VI. Comparisons of Subjective MOS and Objective Quality Estimation Algorithms

A. *Introduction to Objective Quality Measurement*

As part of our subjective quality evaluation, we have also compared various objective perceptual quality estimation algorithms [16], [14], [18], [29], [30]. Objective quality measurement is unbiased by difference of human ears and the listener's own understanding of quality. Therefore the results are always repeatable. The algorithms are usually based on an approximation to the human ear's internal psycho-acoustical representation of an audio signal. As an approximation, these algorithms will inevitably introduce some deviation from subjective quality perceived by the human ear. We present here a comparison of these algorithms versus our subjective MOS results. If the algorithm outputs an objective MOS, we compute its correlation with subjective MOS. Otherwise, the algorithm outputs an objective perceptual distance, and we plot it against subjective MOS. Overall, PESQ performs best, whereas PSQM/PSQM+ and EMBSD seem to have the largest variance in predicting the true MOS. MNB predictions appear to "saturate" for one of our test sets, that is, it almost always thinks the quality is high.

PSQM [14] is originally designed to evaluate codec quality. PSQM+ [18] is an enhancement of PSQM to cover short duration temporal clipping as often seen in wireless communications. PESQ [16] is an intended replacement of PSQM. It is designed to consider artifacts in a voice over IP environment, namely packet loss concealment and playout delay variation. MNB [29] is another objective quality measurement algorithm, defined in the appendix of P.861. EMBSD is an enhancement of MBSD [30].

All these algorithms work by comparing a reference signal and a degraded signal, and output some measure of quality. PESQ directly produces an objective MOS in the range of [1.0,4.5]. The upper bound of 4.5 is probably due to the fact listeners rarely gives scores higher than 4.5 on average. MNB produces 2 perceptual distances (MNB1 & MNB2), and they are mapped into a logistics value with a range of [0,1]. Also known as auditory distance, a larger perceptual distance implies a stronger degradation between the reference (good) signal and degraded signal, and hence lower MOS, but its value is not directly comparable between different algorithms. MNB's logistics value is claimed to have a near linear relationship with subjective MOS, and since PESQ uses a MOS interval of 3.5, we will use (MNB_logistics * 3.5 + 1) to transform it into an objective MOS.

PSQM/PSQM+ and EMBSD output only a perceptual distance, which can be transformed into an objective MOS, but we are not aware of any pre-determined transformation function and tuning parameters for these algorithms. If we perform a test set specific optimization on such transformation, it would be an unfair comparison to PESQ and MNB. Rather, we will present its scatter plot against subjective MOS.

B. Correlation Evaluation

To quantitatively measure the goodness of fit between objective and subjective MOS, we compute the Pearson correlation coefficient ρ for MNB and PESQ. We use two approaches, first, we use the original audio clips (PCM linear-16) as the reference signals to the objective algorithms, then compute the correlation, denoted as ρ_{l16}. Next, we use its corresponding G.729 coded clips without loss as the reference signals, then compute the ρ, denoted as ρ_{g729}. With the second approach, the algorithms are only evaluating degradation only due to packet losses, instead of mixing the codec's quality into the final score. But such usage has not been cited before, so our result could serve as its initial test. The following table shows the correlation results.

Algorithm	Test Set 1		Test Set 2	
	ρ_{l16}	ρ_{g729}	ρ_{l16}	ρ_{g729}
MNB1	0.897	0.885	0.767	0.798
MNB2	0.910	0.935	0.844	0.870
PESQ	0.888	0.902	0.892	0.910

TABLE I
CORRELATIONS OF SUBJECTIVE AND OBJECTIVE MOS

In both test sets, the correlation between objective and subjective MOS increases when the reference input becomes G.729 coded non-loss signal. In test set 1, MNB actually has the highest correlation. However, the correlation value does not tell the whole story. Figure 10(a) shows the visual correlation of objective versus subjective MOS for test set 1. The figure is similar for test set 2, so we will not repeat it here. In the right column, the scatter plot of MNB1 and MNB2 are not linear any more. It appears to have a "saturation" effect at the higher end of MOS (around 4.5), which means MNB (especially MNB1) will almost always think the clip quality is good, whereas the real quality may be much worse. A highly saturated objective MOS does not do any good in terms of MOS prediction. It will over-predict MOS, which is more harmful than no prediction. In this case, the higher correlation coefficient does not imply a linear correlation. On the other hand, PESQ still maintains a linear trend, therefore PESQ produces more predictive objective MOS, although we must notice the overall correlation is still around 0.9, which is not very high.

Figure 10(b) shows the scatter plot of perceptual distance versus subjective MOS. EMBSD and PSQM+ appears to have the largest spread, implying that a relative low degree correlation if it were to be transformed into an objective MOS.

C. Relative MOS Consistency

In both test sets, for most of the data, the output of objective algorithms agrees with subjective MOS in a relative sense. That is, within the same p_u, if MOS(A) $>$ MOS(B), then the objective algorithms agree with this inequality tests more than 80% of the time. Changing the reference signal to G.729 has almost no effect on how the objective algorithms agree with these inequality tests.

From these results, it appears these algorithms are useful in judging the relative audio quality.

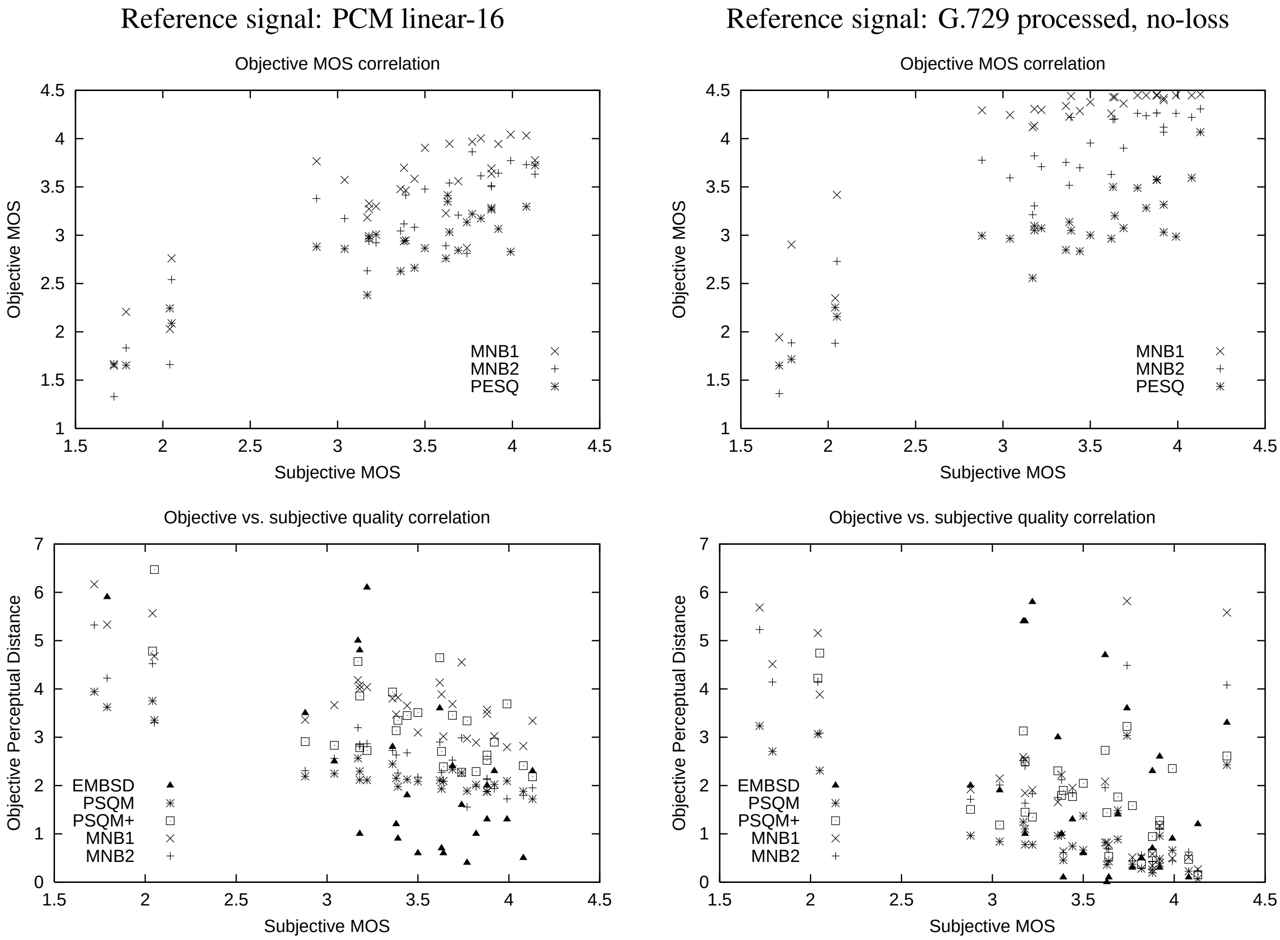
VII. RELATED WORK

Some related studies on perceived quality under bursty have been performed by the telecommunications sector such as ITU, T1A1 [8]. But there are a few limitations which we intend to overcome in this paper.

First, they usually assume an environment more like circuit-switching. For example, packet interval T is usually the same as codec frame length (10 ms for many codecs). Such a small packet size is too inefficient for general use in WAN IP networks, since a IP/UDP/RTP header costs 40 bytes already. With a larger T, loss concealment behaviors could also be quite different.

Second, although a distinction is made between bursty losses and Bernoulli losses, but do not directly quantitatively specify the burstiness, for instance, in term of p_c. This makes the result less comparable unless more information is available.

Third, quality measures such as FEC or low bit-rate redundancy are not evaluated. Bolot [4] studies optimization of low bit-rate redundancy, but the resulting perceived quality is not studied except using a pre-determined utility function that translates bit-rates into a utiltily value



(a) Objective vs. Subjective MOS (b) Perceptual Distance vs. Subjective MOS

Fig. 10. Objective/Subjective correlation: Test Set 1

(quality). However, as we have found in our tests, it generally has lower quality than FEC, and often even worse than default loss concealment. This is true for G.729 with DoD LPC-10 (MOS 2.4-2.5), and informally the author has found G.729 with DoD CELP (MOS 3.2) share the similar distortion problem.

A codec's loss concealment behavior is not always the same. Rosenberg [23] has shown that G.729 can conceal a single frame loss well, but double and triple frame losses degrades its effectiveness significantly. Sanneck [25] and Sun [28] has found that most CELP type codecs can conceal a frame loss well if the loss location is at either an unvoiced or middle/end of a voiced frame. It is much less effective if the loss is at an unvoiced to voiced transition. The decoder state drift is most significant in such cases. In our tests, we have a few occasions where MOS values are unexpected, which we think are due to dependence on loss locations. But so far we are not aware of any algorithm that can reliably use this information to predict MOS values.

In this paper, we only use a normal FEC for loss recovery. However, it is possible to combine our results with speech properties as shown in [25]. An application can then decide on an optimal packet interval T and allocate more FEC data to frames at unvoiced to voiced transitions.

Finally, an alternative to FEC is to have strong loss concealment. For instance, recently an enhanced G.711 codec (http://www.globalipsound.com/technology.html) claims to maintain good quality even at 30% losses. It will greatly affect the concept of Quality of Service if the same can be done for a low bit-rate, frame-based codec.

VIII. SUMMARY

We examine perceived quality of packet audio under bursty loss conditions, in particular how they affect quality improvement measures including FEC and low bit-rate redundancy. We present a formula for re-calibrating the conditional loss probability (p_c) in a Gilbert model when the packet interval T changes. Both the analytical formula and the subjective listening tests confirm that FEC quality is noticeably better when operating at a larger T. But subjective tests also indicate that T should not be too high to prevent a single packet loss penalty to become too high, for instance, that would wipe out an entire piece of speech content (e.g., a phoneme).

We also evaluate perceived quality of an alternative recovery scheme: low bit-rate redundancy. Our test results indicate that its quality is generally lower than that of FEC with a similar overhead. This is based on G.729 as the main codec with DoD LPC-10 as the redundant codec. The lower quality is informally determined to be caused by unnatural transitions during packet losses, possibly due to the lower MOS (quality) of LPC-10. However, an informal test with a higher MOS codec, the DoD CELP, also has similar audible distortions.

Finally, we find PESQ best in predicting objective MOS, but its precision still cannot substitute MOS tests.

REFERENCES

[1] Alan D. Baddeley. *Human memory : theory and practice*. Allyn and Bacon, 1998.

[2] John C. Bellamy. *Digital Telephony*. John Wiley & Sons, Inc., third edition, 2000.

[3] Jean-Chrysostome Bolot. Characterizing end-to-end packet delay and loss in the Internet. *Journal of High Speed Networks*, 2(3):305–323, 1993.

[4] Jean-Chrysostome Bolot, Sacha Fosse-Parisis, and Don Towsley. Adaptive FEC-Based error control for interactive audio in the Internet. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, New York, March 1999.

[5] Alan Clark. Modeling the effects of burst packet loss and recency on subjective voice quality. In *Internet Telephony Workshop 2001*, New York, April 2001.

[6] Pascal Frossard. Fec performance in multimedia streaming. *IEEE Communications Letters*, 5(3):122–124, March 2001.

[7] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. 1993.

[8] Working group T1A1.7. Results of a subjective listening test for g.711 with frame erasure concealment. Technical report, Committee T1, May 1999.

[9] Vicky Hardman, Angela Sasse, Mark Handley, and Anna Watson. Reliable audio for use over the Internet. In *Proc. of INET'95*, Honolulu, Hawaii, June 1995.

[10] International Telecommunication Union. Objective measurement of active speech level. Recommendation P.56, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, March 1993.

[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, March 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, March 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, February 1996.

[14] International Telecommunication Union. Objective quality measurement of telephone-band (300-3400 hz) speech codecs. Recommendation P.861, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, February 1998.

[15] International Telecommunication Union. Pulse code modulation (PCM) of voice frequencies. Recommendation G.711, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, November 1998.

[16] International Telecommunication Union. Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Recommendation P.862, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, February 2001.

[17] Wenyu Jiang and Henning Schulzrinne. Modeling of packet loss and delay and their effect on real-time multimedia service quality. In *Proc. International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, June 2000.

[18] the Netherlands KPN. COM 12-20-E, improvement of the p.861 perceptual speech quality measure. Contributions, Telecommunication Standardization Sector of ITU, 1997.

[19] Vern Paxson. End-to-end internet packet dynamics. In *SIGCOMM Symposium on Communications Architectures and Protocols*, Cannes, France, September 1997.

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

[21] Colin Perkins, Orion Hodson, and Vicky Hardman. A survey of packet loss recovery techniques for streaming audio. *IEEE Network*, 12(5):40–48, September 1998.

[22] Jonathan Rosenberg, Lili Qiu, and Henning Schulzrinne. 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, March 2000.

[23] Jonathan D. Rosenberg. G.729 error recovery for internet telephony. Technical report, Columbia University, 1997.

[24] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, H. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, AU-17(3):225–238, September 1969.

[25] Henning Sanneck and Nguyen Tuong Long Le. Speech property-based FEC for Internet Telephony applications. In *Proceedings of the SPIE/ACM SIGMM Multimedia Computing and Networking Conference (MMCN)*, pages 38–51, San Jose, California, January 2000.

[26] Henning Schulzrinne, James F. Kurose, and Donald Towsley. Loss correlation for queues with bursty input streams. In *Conference Record of the International Conference on Communications (ICC)*, volume 1, pages 0219–0224 (308.4), Chicago, Illinois, June 1992. IEEE.

[27] Henning Schulzrinne and Jonathan Rosenberg. Internet telephony: Architecture and protocols – an IETF perspective. *Computer Networks and ISDN Systems*, 31(3):237–255, February 1999.

[28] Lingfen Sun, Graham Wade, Benn Lines, and Emmanuel Ifeachor. Impact of packet loss location on perceived speech quality. In *Internet Telephony Workshop 2001*, New York, April 2001.

[29] Stephen Voran. Objective estimation of perceived speech quality, part i: development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing*, 1999.

[30] Wonho Yang and R. Yantorno. Improvement of mbsd by scaling noise masking threshold and correlation analysis with mos difference instead of mos. In *IEEE ICASSP*, pages 673–676, 1999.