Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications

DAVID J. GOODMAN, MEMBER, IEEE, GORDON B. LOCKHART, ONDRIA J. WASEM, STUDENT MEMBER, IEEE, AND WAI-CHOONG WONG

Abstract—**Packet communication systems cannot, in general, guarantee accurate and prompt delivery of every packet. The effect of network congestion and transmission impairments on data packets is extended delay; in voice communications these problems lead to lost packets.**

When some speech packets are not available, the simplest response of a receiving terminal is to substitute silence for the missing speech. Here, we explore techniques for replacing missing speech with waveform segments from correctly received packets in order to increase the maximum tolerable missing packet rate.

After presenting a simple formula for predicting the probability of waveform substitution failure as a function of packet duration and packet loss rate, we introduce two techniques for selecting substitution waveforms. One method is based on pattern matching and the other technique explicitly estimates voicing and pitch. Both approaches achieve substantial improvements in speech quality relative to silence substitution.

After waveform substitution, a significant component of the perceived distortion is due to discontinuities at packet boundaries. To reduce this distortion, we introduce a simple smoothing procedure.

I. INTRODUCTION

PACKET speech communication may play an important role in the evolution of combined voice and data services. Although the advantages of communicating computer data in packets are well documented [1], the editors of a recent collection of papers report that "the jury is still out" on the merits of packet speech [2]. In contrast to packet data transmission where delays are allowed to build up as traffic increases, speech communication requires prompt packet delivery. Beyond some time limit, delayed speech packets are useless at the receiving terminal and are discarded by the system. Packet loss, therefore, has a major effect on speech quality and the consequent constraints on packet dropping rates affect system costs. In formal listening tests conducted to assess the effects of missing packets on speech quality [3], [4], silent gaps replaced the missing speech packets, and it was determined that packet loss rates up to about 1 percent were tolerable. There are reports of other techniques for dealing with lost packets, such as repeating previous packets or, if the speech has been processed by a vocoder, synthesizing new speech from previously received analysis data [5]. Another approach is to construct speech packets at the transmitter in a manner that facilitates the recovery of lost packets. Jayant and Christensen have experimented with an interleaving technique that places odd-numbered samples and even-numbered samples in different packets [6]. When an isolated packet is missing at the receiver, the samples in a neighboring packet are used to estimate the missing samples.

By contrast, the packet reconstruction techniques presented in this paper operate on conventional PCM packets containing consecutive speech samples. They apply only at the receiver and incur negligible processing delay. We assume that the receiver learns the positions of missing packets from time stamps and/or sequence numbers in the headers of correctly received packets [5]. Not only does this header information guide the replacement process, but it also allows the receiving terminal to establish correct timing. Timing uncertainties arise from variable transmission delay and from speech activity detection which suppresses all packet transmission during silent intervals.

The simplest reconstruction technique is merely to set all samples to zero when packets are missing and to accept the distortion caused by the resulting gaps in received speech. This "zero substitution" may be acceptable for very small probabilities of packet loss, but for rates greater than about 1 percent there is much to be gained by attempting to reconstruct the waveform of the missing packet.

The term "waveform substitution" refers to reconstruction of missing packets by substitution of past waveform segments. Because it is likely that the contents of a missing packet will resemble immediately preceding speech, this approach is attractive in that waveform synthesis is not required. Instead, the substitute waveform is selected from speech already available at the receiver. Packet reconstruction thus amounts to selecting a speech segment and placing it in the missing packet time slot.

Manuscript received November 30, 1985; revised March 22, 1986.

D. J. Goodman is with the AT&T Bell Laboratories, Crawford Hill Laboratory, Holmdel, NJ 07733.

G. B. Lockhart is with the University of Leeds, Leeds LS2 9TS, England.

O. J. Wasem is with the Massachusetts Institute of Technology, Cambridge, MA 02139.

W.-C. Wong is with the National University of Singapore, Singapore, 0511.

IEEE Log Number 8610377.

0096-3518/86/1200-1440$01.00 © 1986 IEEE

The simplest waveform substitution scheme replaces a missing packet with the previous packet. This scheme, however, introduces distortion due to discontinuities at packet edges, an impairment that is reduced by more intelligent waveform selection techniques which exploit the periodicities in voiced speech.

II. Performance of Waveform Substitution Schemes

Waveform substitution will be effective provided the character of the speech signal does not change significantly during a missing packet. Speech waveforms display quasi-stationary intervals which for the most part fall into one of three distinct categories: high-energy voiced speech with strong pitch periodicity, low-level unvoiced speech, and silence [7]. Significant changes in the character of the speech, and therefore incorrect waveform substitution, can occur for either of the following two reasons:

a) the missing segment is so long that the speech signal is nonstationary, or

b) there is a transition from one category to another within the missing segment.

In the Appendix, we use a simple probability model to describe the effects of these two failure conditions on the performance of an idealized waveform substitution scheme. Our simplifying assumption is that waveform substitution fails when either condition a) or condition b) occurs, and that it is perfect when neither condition occurs. We then analyze the probabilities of these conditions as functions of packet duration T_p ms, and missing packet probability p. We introduce the variables

$$C = 32/T_p, \tag{1}$$

the maximum number of packet intervals over which the speech signal can be considered stationary, and

$$t = \exp(-0.0052T_p) \approx 1 - 0.0052T_p, \tag{2}$$

the probability that there is no category transition during any particular packet. In terms of these variables, the probability of unsuccessful waveform substitution is

$$P_f = 1 - (1-p)\frac{1-(pt)^{C+1}}{1-pt}. \tag{3}$$

Fig. 1 displays this failure probability as a function of missing packet probability, for packet durations T_p = 8, 16, and 32 ms. Observing that $P_f = p$ for zero substitution and assuming that, for example, P_f = 0.01 is a packet communication performance objective [3], we see in Fig. 1 that relative to zero substitution, which can tolerate a missing packet rate of 1 percent, waveform substitution can potentially increase the acceptable missing packet probability to 19 percent, 10 percent, or 5 percent when the packet duration is 8, 16, or 32 ms, respectively.

III. Perceptual Aspects, Packet Merging

While this probability model provides a guide to the quality of waveform substitution, it is also important to consider the perceptual effects of substitution failures. Listening experience suggests that the transition failure b) is perceptually the less damaging of the two mechanisms because it merely truncates or extends the current speech category for a short interval. On the other hand, failure due to mechanism a) is likely to be more disturbing because it sustains for a long interval (greater than 32 ms)—a sound selected from previous speech packets. This leads to a perceptible discontinuity when correct packets reappear, and if the sustained sound is voiced, it is heard as a disturbing buzz or whistle.

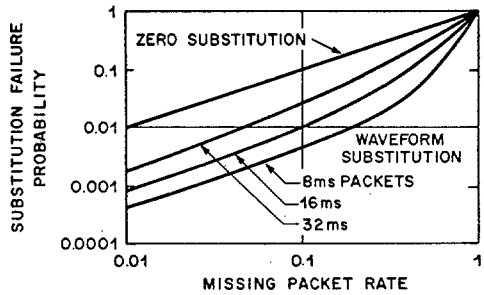


Fig. 1. Substitution failure probability as a function of missing packet rate. Within a 1 percent failure-rate criterion, the maximum missing packet rate is approximately 5, 10, or 20 percent when the packet length is 32 ms, 16 ms, or 8 ms, respectively.

In addition to the two failure mechanisms that we analyzed, we observed perceptible distortions due to small discontinuities at some of the boundaries between correct packets and substitution packets. To reduce the audibility of this distortion, we found it helpful to increase slightly the duration of substitution packets so that packet ends can be "merged" with the corresponding correct packets. Our merging procedure consists of raised-cosine weighting and addition of overlapping packet segments as indicated in Fig. 2. The merge duration T_m = 1 ms produces acceptable results and, in fact, this merging also makes zero substitution less objectonable even though it alters the first or last millisecond of some correct packets.

IV. Waveform Substitution Based on Pattern Matching

A. Overview

Figs. 3–5 illustrate this approach. In Fig. 3 we see a speech waveform divided into packets each containing L samples. One of the packets is missing and the algorithm searches previous packets to find L samples that resemble the missing packet. To do so, it uses as a *template* the M speech samples that came just before the missing packet. Fig. 4 indicates that the algorithm scans a *search window* of duration N samples to find the M samples that best match the template. It then uses as a replacement packet the L samples that follow the best match. Fig. 5 shows the reconstructed waveform.

To apply the merging technique of Fig. 2, we move the search window (in Fig. 3) T_m ms to the left and append T_m ms of received speech to each end of the replacement packet. The raised-cosine smoothing is then applied to the

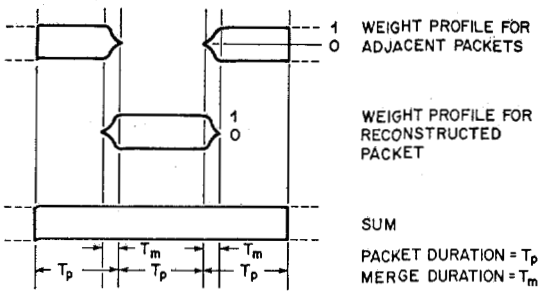


Fig. 2. Raised cosine weight profiles for merging a substitution packet into received speech.

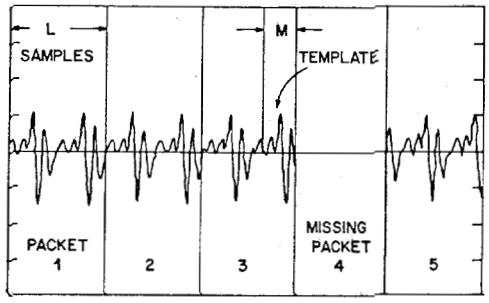


Fig. 3. Speech waveform divided into five packets, each with L samples. Packet 4 is missing, and the final M samples of packet 3 comprise a template to be used in a search for a substitution packet.

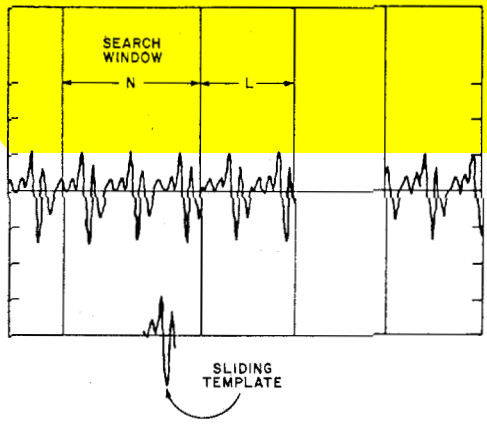


Fig. 4. The template slides along a search window containing N samples to find M samples that best match the template.

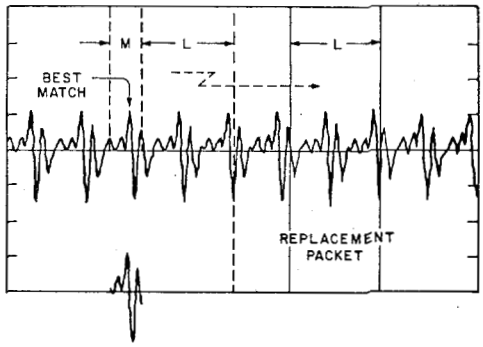


Fig. 5. The substitution packet contains the L samples immediately following the best match to the template.

replacement packet and to the packets that precede and follow it.

B. Pattern Matching

There are several methods of pattern matching which yield almost equivalent results. One method is cross correlation of samples in the template and samples of the search window. If the samples of the template are $x(i)$, and the samples of the search window are $y(i)$, then the cross-correlation formula is

$$C(n) = \frac{\sum_{m=1}^{M} x(m)\, y(n+m)}{\sum_{m=1}^{M} [y(n+m)]^2}, \quad n = 1, 2, \cdots, N, \tag{4}$$

where M is the number of samples in the template and n identifies the position of the template as it slides along the search window. The result of the search is the value of n corresponding to the maximum $C(n)$.

A simplified version of $C(n)$, which is attractive in some practical implementations, is the sign correlation,

$$S(n) = \sum_{m=1}^{M} \operatorname{sgn}[x(m)] \operatorname{sgn}[y(n+m)] \tag{5}$$

where sgn (x) is $+1$ when $x > 0$ and -1 when $x < 0$.

Another approach to pattern matching is based on waveform differences. As the template slides along the search window, the algorithm seeks the minimum sum of absolute differences. In order that the result be sensitive to waveform shapes rather than level changes, the speech segments are normalized first. We have considered three methods of normalization. One is to divide the samples of each segment by the square root of the energy of that segment. This leads to the difference measure

$$D_1(n) = \sum_{m=1}^{M} \left| \frac{x(m)}{\sqrt{\sum_{j=1}^{M} [x(j)]^2}} - \frac{y(n+m)}{\sqrt{\sum_{j=1}^{M} [y(n+j)]^2}} \right|. \tag{6}$$

Note that the first denominator is independent of n for a given missing packet, because the template never changes. However, the second denominator changes each time the template advances one sample. Another way to normalize is by the sum of the absolute magnitudes of the samples

$$D_2(n) = \sum_{m=1}^{M} \left| \frac{x(m)}{\sum_{j=1}^{M} |x(j)|} - \frac{y(n+m)}{\sum_{j=1}^{M} |y(n+j)|} \right|. \tag{7}$$

A third way is to divide by the peak-to-peak amplitude of the segment. In the case of the search window, the peak-to-peak amplitude is not across the whole window, but only across the segment being compared to the template. The formula for the resulting difference measure is:

$$D_3(n) = \sum_{m=1}^{M} \left| \frac{x(m)}{x_{\max} - x_{\min}} - \frac{y(n+m)}{y_{\max} - y_{\min}} \right|, \quad (8)$$

where

$$x_{\max} = \max [x(1), x(2), \cdots, x(M)],$$

$$y_{\max} = \max [y(n+1), y(n+2), \cdots, y(n+M)]$$

and similarly for $x_{\min}$ and $y_{\min}$.

In our simulation experiments with the five pattern-matching measures, we judged the difference methods to produce speech that sounds slightly better than speech produced with the correlation measures and that the normalizations of (6) and (7) are superior to (8).

To improve the perceived quality of the reconstructed speech, we found it helpful to adjust the amplitude of the substitution packet to match that of the preceding packet. To achieve this we experimented with three measures of packet amplitude, root-mean-square, mean-absolute-value, and peak-to-peak, analogous to the normalizations in (6), (7), and (8), respectively. This amplitude adjustment has a stronger influence on quality than the choice of a pattern-matching criterion. The root-mean-square and mean-absolute-value measures are comparable to each other and superior to the peak-to-peak measure.

C. Two-Sided Approach

We experimented with an extension of the pattern-matching algorithm that uses speech received after the missing packet in addition to speech that precedes it. This two-sided scheme selects a "future" replacement packet in a manner that is exactly analogous to the past-packet selection of Figs. 3–5. The final replacement packet is a weighted sum of past and future selections.

This scheme is clearly more complicated than the one-sided approach and it incurs the performance penalty of added delay. Formal listening tests will be undertaken to assess the extent to which the added complexity and delay of the two-sided scheme are compensated by improved quality.

V. Waveform Substitution Based on Pitch Detection

Our other approach to waveform selection is based on pitch detection. If a reliable estimate of pitch period P ms is available for a voiced segment, then a missing packet can be reconstructed by repetitions of the last P ms of available speech in the missing packet time slot. Such a strategy aims at continuity on the basis of the most recent waveform information but makes no allowance for changes that may occur during missing packets.

A wide variety of methods is available for speech pitch detection. Our method can be interpreted as a variant of a parallel processing method proposed by Gold and Rabiner [8]. Two parallel detectors are employed which continually detect positive and negative peaks, respectively, of the speech signal. Center-clipping [7] with threshold CT provides a crude voiced/unvoiced classification and each peak detector attempts to isolate and identify a single "significant" peak in each pitch period. The operation of the positive peak detector is specified in Fig. 6.

At the beginning of its cycle, the positive peak detector updates the value of MAX with successive local maxima of speech samples until no update has occurred for HLD samples. The time position of the last update is then stored as a significant peak and MAX is allowed to decay exponentially until exceeded by a speech sample whereupon the cycle restarts. HLD should be greater than the number of samples corresponding to the smallest expected pitch period. The decay should be fast enough to accommodate peak detection in a succession of pitch periods with decreasing amplitude but slow enough to reject spurious low-level peaks in pitch periods of long duration.

The negative peak detector operates in a similar way for negative peaks and both detectors store the time positions of the latest 3 significant peaks. If a missing packet occurs, then a reconstructed packet is generated according to one of the following strategies. In all cases the final reconstructed packet is merged with adjacent packets as described in Section III.

1) If the last significant peak to be detected by both detectors occurred more than a constant T_r ms before the beginning of the missing packet, then it is assumed that the previous packet is unvoiced and a copy of the previous packet is taken as the reconstructed packet.

2) If 1) above is not satisfied, then two pitch period estimates are calculated for each pitch detector from the stored time positions of the latest three positive and negative significant peaks. Any estimate which falls outside an expected range from P_L to P_H is discarded. A "confident" estimate P is made only if agreement is secured within 8 percent between, i) two pitch estimates from the same detector, and/or ii) the latest estimates from both detectors. If i) is true for the estimates from only one detector, then P is taken as the average of the two estimates. However, if both detectors provide apparently reliable but contradictory estimates, then the higher estimate is accepted on the assumption that the other detector has erroneously detected more than one significant peak per pitch period.

3) If conditions 1) or 2) above are not met, it is assumed that pitch detection has failed in the presence of voiced speech and a copy of the previous packet is taken as the reconstructed packet.

VI. Experimental Results

In order to evaluate the effects of substitution parameters and the missing packet rate on speech quality, we have simulated the transmission of 11.2 s of speech sampled at 8 kHz. The speech material consists of one sentence from each of four speakers, two men and two women. Although, in practical networks, the missing packet statistics are likely to be bursty, the missing packets in our experiments were randomly distributed in time. We believe that our conclusions about the *relative* effects

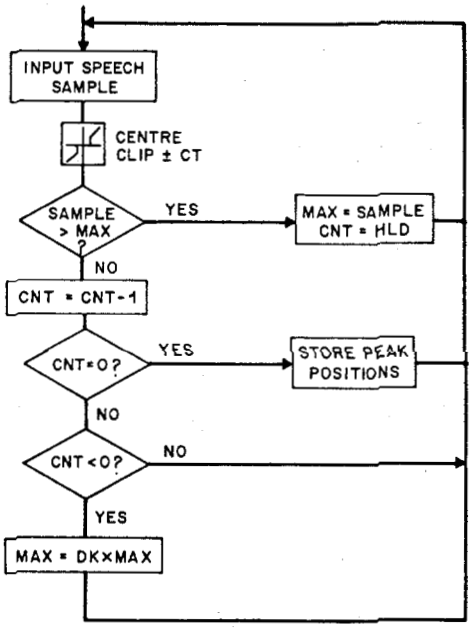


Fig. 6. Flowchart of the pitch estimation procedure, showing the procedure for determining the number of samples between positive waveform peaks.

of various parameter settings are generally applicable. However, the *overall* effect of waveform substitution has to be assessed in the presence of packet loss patterns that are representative of each potential application.

A. Pattern Matching

In Section IV-B, we defined several matching criteria and discussed their influence on the quality of the reconstruction scheme. We now present the effects of packet length (L), template length (M), and search window length (N) on signal-to-noise ratio and on our perceptions of speech quality. In performing these evaluations, we have used the magnitude difference measure (7) for pattern matching and the root-mean-square amplitude adjustment of replacement packets.

1) Packet Size: For a given fraction of packets missing, the packet size has a strong effect on the perceived nature of the reconstructed speech. With very small packets (1 or 2 ms, $L = 8$ or 16), with 10 percent of the packets missing, there is a constant, annoying crackle. For very large packets (32 ms or more), the speech sounds as though the person is trying to gargle while speaking. For sizes in between, the crackles become pops, and occur infrequently, rather than constantly like the crackle. To our ears, the packet size most tolerant to packet loss is 8 ms (64 samples), although 16 ms is also good. This observation conforms approximately to the reports of earlier researchers who found that with silence substitution 16–32 ms packets were most tolerant of packet loss [6]. Signal-to-noise ratio of reconstructed speech is not a good indication of how the quality changes with packet size; it improves as the packets get smaller and smaller.

2) *Search Window:* There is an optimum search window duration. If the search window is too short, it omits the best reconstruction waveform. If it is too long, it contains speech that is unrelated to the missing packet. Nevertheless, there is a chance that a small segment of this speech is well matched to the M samples in the template, a situation which can result in the selection of a suboptimum reconstruction segment. In our experiments, we found that regardless of packet size, a 16 ms search window was best for the basic (one-sided) pattern-matching scheme. With the two-sided technique of Section IV-C, 8 ms was the best search window duration. Speech quality does not deteriorate appreciably when the search window is longer than optimum.

We observed that signal-to-noise ratio is a reasonably good indicator of the relationship of perceived quality to search window duration. Fig. 7 shows the total SNR (measured across 11 s of speech) as a function of the number of samples in the search window. Fig. 8 displays the average SNR (in dB) per missing packet, a measure found in a previous study to be a good indicator of the relative quality of reconstruction methods [6]. The measurements in Figs. 7 and 8 were obtained with $L = 64$ samples per packet and a missing packet rate of 9.2 percent. For the basic method, the template contains $M = 32$ samples; for the two-side scheme, $M = 16$.

Note in Figs. 7 and 8 that the shortest search window is equal in duration to the template. In this case, the algorithm replaces the missing packet with the previous packet.

3) Template: As in the case of the search window, there is a minimum acceptable template duration. If M is too small, there is simply insufficient speech information. The quality also goes down if the template is too long. Again, the best sizes, 4 ms ($M = 32$) for the basic scheme and 2 ms ($M = 16$) for the two-sided scheme, appear to be independent of the packet size.

SNR correlates well with the effect of template duration on perceived quality. Fig. 9 shows the average SNR per missing packet as a function of template size.

4) Missing Packet Rate: As expected, the SNR and the perceived quality decline as the fraction of packets missing increases. Fig. 10 shows the dependence of total SNR on missing packet rate for the two pattern-matching methods and for missing packets replaced by silent gaps. Our listening experience suggests that communication breaks down when more than 30 percent of the speech is lost. When only half of the packets arrive, the replacement algorithm produces speech interspersed with beeps and chirps very similar to the voice of the robot R2D2 in the movie "Star Wars." This is because when many packets in a row are lost, the one-sided scheme repeats the same segment, creating a highly periodic signal.

B. Pitch Detection Approach

Table I lists the parameters used in this simulation. We found that in order to accommodate pitch periods ranging from 2.5 to 12.5 ms, some degree of parameter adaptation was necessary in the pitch detector. The algorithm illus-

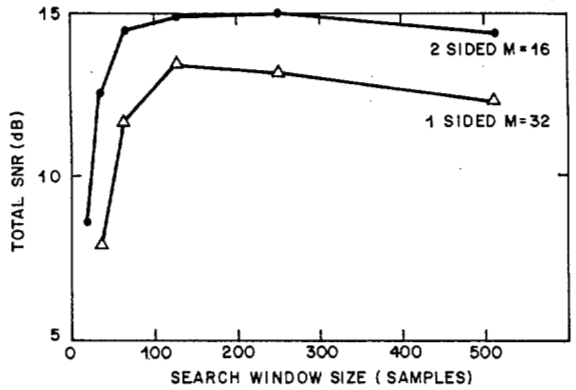


Fig. 7. Total signal-to-noise ratio as a function of search window size. There are 64 samples per packet and 9.2 percent of the packets are missing. In the basic (one-sided) version, $M = 16$ samples per correlation segment; $M = 32$ with two-sided reconstruction.

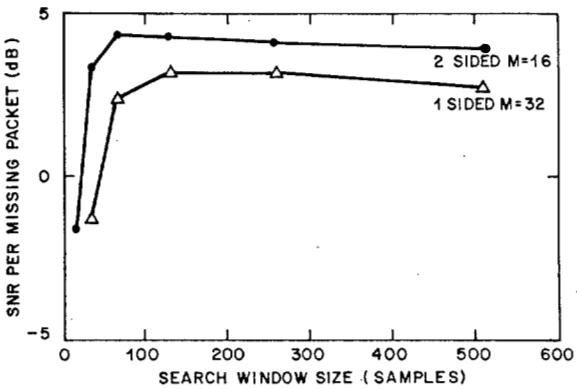


Fig. 8. Average signal-to-noise ratio per missing packet for the same condition as Fig. 7.

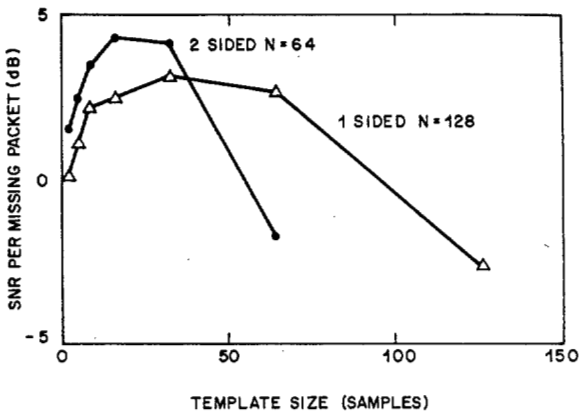


Fig. 9. Average signal-to-noise ratio per missing packet as a function of correlation window size. There are 64 samples per packet and 9.2 percent of the packets are missing. With one-sided estimation, there are $N = 128$ samples in the search window; $N = 64$ in the two-sided case.

trated in Fig. 6 was therefore modified to allow the hold parameter *HLD* to increase, from an initial value of 20 samples, by one-quarter of the number of samples taken during the decay phase of the previous detection cycle. The decay factor *DK* was also made partially adaptive by setting $DK = 1\text{-}0.6/HLD$ at the beginning of a cycle to achieve some lengthening of the decay time constant for longer values of pitch period.

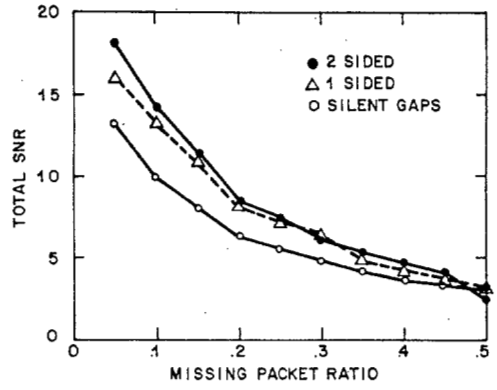


Fig. 10. Total SNR as a function of missing packet rate for both versions of the pattern-matching scheme and for silence substitution.

TABLE I
PARAMETER VALUES USED IN SIMULATION OF PITCH DETECTION ALGORITHM

PCM sampling rate	= 8 kHz
Packet duration T_p	= 16 ms
Merge duration T_m	= 1 ms
Voiced/unvoiced decision threshold CT	= 10 percent of peak speech
Maximum allowance for acceptance of signal peaks T_r	= 16 ms
Minimum acceptable pitch period P_L	= 2.5 ms
Maximum acceptable pitch period P_H	= 12.5 ms

The relative frequencies of the three decision strategies (Section V) invoked for missing packet reconstruction are listed in Table II for missing packet probabilities of 0.1, 0.2, 0.3, and 0.4. In order to reduce dependence on specific patterns of missing packets, the entries in Table II are averages over 3 runs using different random number seeds to select missing packets. The entries for strategy 2) are subdivided according to how a confident estimate P of pitch period was derived. Within this category, the estimates are usually made with the confidence of both detectors and examination of the speech records confirms that the measure of periodicity is invariably correct in these cases. Cases where reliable estimates from both detectors disagreed significantly are not shown as they occurred with extremely small frequency. Decisions for strategy 2) taken with the confidence of only one detector were usually considered satisfactory on visual inspection of the waveform records. Such cases occurred typically when high-level spurious peaks of either positive or negative polarity disabled the operation of one detector but not both. A small proportion of decisions was made with only the agreement of the latest estimates from both detectors. These decisions usually occurred in the vicinity of a transition, and provided a rapid response at the beginning of a voiced segment.

Strategy 3), invoked when neither voiced nor unvoiced decisions can be made, was used for about 8 percent of missing packets, usually in the vicinity of transitions where voiced/unvoiced distinctions become difficult to make.

TABLE II
PITCH DETECTION ALGORITHM PERFORMANCE

	Missing Packet Probability			
	0.1	0.2	0.3	0.4
No. of missing packets	69	139	209	279
Strategy(1): Unvoiced (percent)	41.1	45.1	42.9	45.6
Strategy(2): Voiced (percent)				
Both detectors agree	23.2	32.4	28.1	31.3
Only +ve detector confident	12.6	7.7	10.0	7.4
Only −ve detector confident	10.6	6.2	4.9	4.2
Both detectors confident only on latest estimates	4.4	2.4	3.7	3.7
Strategy(3) Voicing ambiguous (percent)	8.2	6.2	10.4	7.8
Total signal/distortion (dB)	10.6	8.00	5.48	4.04
Average signal/distortion per missing packet (dB)	1.25	1.28	0.53	−0.01

Table II also displays SNR measured over the entire 11 s speech sample and the SNR per missing packet. Both measures tend to decrease with increasing missing packet probability. The distortion per missing packet increases because the success of packet reconstruction becomes limited by the frequent occurrence of long missing packet chains [failure condition a) in Section II]. While the total SNR's in Table II are comparable to those displayed in Fig. 10 for the pattern-matching schemes, it is impossible, owing to the loose connection between SNR and subjective quality, to deduce the relative merits of the two approaches from these data.

VII. IMPLEMENTATION COMPLEXITY

Essentially two computational processes are required for the implementation of a waveform substitution scheme. The first involves extraction of the necessary speech parameters such as pitch period or pattern-matching location while the second involves waveform placement in the missing packet time slot on the basis of available parameters. Although the latter may be a logically complicated decision process in the pitch estimation algorithm involving strategies such as 1), 2), and 3) in Section V, it is required no more than once per packet duration, and therefore, the computational demand averaged over this period is likely to be dominated by the parameter extraction process.

We programmed the WE DSP 20® Digital Signal Processor [9] to implement a simplified version of the pitch detection scheme operating with 16 ms packets of μ-encoded PCM at an 8 kHz sampling rate. On the basis of the experience gained, it is possible to give an indication of the number of DSP 20 instructions required for some waveform substitution schemes. The entries in Table III refer to executed DSP 20 instructions per speech sample averaged over one packet and for purposes of comparison include the trivial case of direct substitution (of zeros or the last packet). "Pure" autocorrelation refers to pitch detection by computation of the complete discrete autocorrelation function for 2.5–12.5 ms delays over a packet-length window. "Fast autocorrelation" refers to the same process using computational shortcuts [7] which might reasonably reduce computational effort by a factor of two. Although the pattern-matching method appears more complex as a software simulation, it requires about the same computational effort as the pitch detection method since the DSP 20 favors pipelined arithmetic rather than the conditional branching required for the pitch detection illustrated in Fig. 6.

TABLE III
ESTIMATED DSP INSTRUCTIONS PER SPEECH SAMPLE FOR VARIOUS RECONSTRUCTION SCHEMES

Pure Autocorrelation	100
Pattern Matching	70
Pitch Detection	70
Fast Autocorrelation	50
Direct Substitution	3

VIII. CONCLUSION

Waveform substitution schemes form a class of straightforward methods for the reconstruction of missing voice packets. We have argued that the performance of such schemes is governed by essentially two characteristics of speech. The first is the maximum period over which speech can be considered stationary, and the second concerns the transitions which occur between voiced and unvoiced or silent speech segments. We have derived an expression for the probability of failure of an idealized substitution scheme and presented two practical schemes one based on pattern matching and the other based on pitch detection. Although it is difficult to make a quantitative assessment of the quality of these schemes in relation to the ideal, simulation results appear generally to conform to our performance predictions.

In addition to the measurements (Figs. 7–10 and Table II) obtained in simulations of 11.2 s of speech transmission, we have gained considerable experience listening to simulations with other source speech and to the real-time implementations of some of the packet replacement schemes. Our perceptions of speech quality are generally consistent with the data, and it is our impression that in-

telligible speech can be obtained in the face of missing packet probabilities up to about 0.3. However, we have also learned that subjective quality is very sensitive to specific patterns of missing packets, particularly chance coincidences of long missing packet chains with high-level voiced speech. For example, with 700 16-ms packets in the 11.2 s test signal, only about one missing packet chain of 3 or more packets can be expected with a missing packet probability of 0.1. The perceptual effect of failed reconstruction within the resulting 48 ms segment will therefore depend critically on the time position of the segment and distortion may vary from negligible during silent periods to gross during high-level voiced speech. Because of this phenomenon, very careful consideration should be given to the formulation of subjective tests for waveform substitution schemes.

Appendix
Waveform Substitution Failure Probability

We derive (3) which is the probability of occurrence of failure condition a) or condition b) (or both simultaneously) defined in Section II. We begin by observing that condition a) occurs only when the total duration of a sequence of missing packets becomes significant relative to expected changes in long-term speech parameters such as envelope and pitch. Since speech parameters can be assumed stationary for no more than about 30 ms [6], the number of contiguous missing packets that can be tolerated before condition a) occurs in approximately $30/T_p$, where T_p ms is the packet duration. Because packet durations of 8, 16, and 32 ms are of practical interest, we adopt (1) for C, the maximum number of consecutive missing packets consistent with successful waveform substitution. These durations correspond to $C = 4$, 2, and 1 consecutive missing packets.

The second failure condition b) occurs when there is a category transition within the duration of the missing packet. In this case, all available substitution waveforms (from previous packets) will be incorrect after the transition. Examining several speech waveforms, we estimate that in English there are 5.2 category transitions per second. Assuming their occurrences conform to a Poisson probability model, we have t, in (2), the probability of no transition over the duration of a packet lasting T_p ms. If there is a sequence of r consecutive missing packets, the waveform substitution fails if $r > C$, which means that regardless of transitions, the duration of the missing speech is too long to allow accurate substitution from past waveform segments. If $r \leq C$, the substitution will fail if there is a transition somewhere in the sequence of r missing packets. Assuming that transitions occur independently from packet to packet, the probability of no transition in a sequence of r packets is t^r and the probability of failure is

$$P_f(r) = 1 - t^r; \quad r \leq C$$
$$= 1; \quad r > C. \tag{9}$$

If the probability of packet loss is p, independent of packet position, the probability of a sequence of r missing packets is

$$P_m(r) = p^r(1 - p). \tag{10}$$

Thus, the total probability of failure of the waveform substitution method is

$$P_f = \sum_{r=0}^{\infty} P_f(r)\, P_m(r), \tag{11}$$

which, combined with (10) and (9), produces (3).

Acknowledgment

We are grateful for the advice and encouragement of J. Evans, R. Valenzuela, P. McLane, and S. O'Riordan.

References

[1] R. D. Rosner, *Packet Switching Tomorrow's Communications Today*. Belmont, CA: Lifetime Learning Publications, 1982.
[2] M. Decina and D. Vlack, "Voice by the packet?" *IEEE J. Selected Areas Commun.*, vol. SAC-1, pp. 961–962, Dec. 1983.
[3] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. Commun.*, vol. COM-33, pp. 801–808, Aug. 1985.
[4] J. Gruber and N. Le, "Performance requirements for integrated voice/data networks," *IEEE J. Selected Areas Commun.*, vol. SAC-1, pp. 981–1005, Dec. 1983.
[5] C. J. Weinstein and J. W. Forgie, "Experience with speech communication in packet networks," *IEEE J. Selected Areas Commun.*, vol. SAC-1, pp. 963–980, Dec. 1983.
[6] N. S. Jayant and S. W. Christensen, "Effects of packet losses in waveform coded speech and improvements due to an odd-even sample interpolation procedure," *IEEE Trans. Commun.*, vol. COM-29, pp. 101–109, Feb. 1981.
[7] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[8] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pt. 2, pp. 442–448, Aug. 1969.
[9] J. S. Thompson and J. R. Boddie, "An LSI digital signal processor," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1980, pp. 383–385.

David J. Goodman (M'67) received the Bachelor's degree from Rensselaer Polytechnic Institute, Troy, NY, in 1960, the Master's degree from New York University, New York, NY, in 1962, and the Doctorate degree from Imperial College, University of London, England, in 1967, all in electrical engineering.

Since 1967 he has been at AT&T Bell Laboratories, Holmdel, NJ, where he is currently Head of the Communications Methods Research Department. He has done research in several aspects of source coding, digital signal processing, and communications. Recently he and his colleagues have been studying short range communications networks, including mobile radio and indoor wireless communications. Since 1983 he has been a Visiting Professor in the Electrical Engineering Department of Imperial College, University of London.

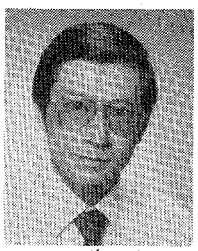
Gordon B. Lockhart was born in Edinburgh, Scotland, on July 13, 1942. He received the B.Sc.(Eng.) (with Honours) and the M.Sc. degrees from Aberdeen University, Aberdeen, U.K., in 1965 and 1966, respectively. He received the Ph.D. degree from Imperial College, University of London, England, in 1970.

He joined the Communications Section of the Department of Electrical Engineering, Imperial College, in 1966 to work on SSB techniques for broadcasting. From 1969 to 1971 he was employed as a Research Fellow at the University of Technology, Loughborough, sponsored by the Joint Speech Research Unit, working on speech encoding—particularly delta-modulation. He has been employed as a Lecturer in the Department of Electrical and Electronic Engineering, University of Leeds, since 1971. In 1984 he spent a 3 month sabbatical period at the Crawford Hill Laboratory, AT&T Bell Laboratories, Holmdel, NJ, as a Visiting Consultant.

Ondria J. Wasem (S'84) was born on March 26, 1964. She received the B.S. and M.S. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in June 1986.

During the Summer of 1983 she worked as an Intern at AT&T Bell Laboratories, Holmdel, NJ, in the Robotics Principles Research Department. During the Summer of 1984 and from June 1985 through January 1986, she worked again as an Intern at AT&T Bell Laboratories, this time in the Communication Methods Research Department. There she completed her Master's thesis on reconstructing missing packets of PCM and ADPCM encoded speech. She is currently working on the Ph.D. degree in electrical engineering and computer science under a National Science Foundation Fellowship at M.I.T.

Mrs. Wasem chaired the M.I.T. Student Chapter of the IEEE and the M.I.T. Department of Electrical Engineering and Computer Science Student Faculty Committee from September 1984 through May 1985. She is also a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.



Wai-Choong Wong received the B.Sc.(Hons.) and Ph.D. degrees, both in electronic and electrical engineering, from the University of Technology, Loughborough, in 1976 and 1980, respectively.

From 1980 to 1983 he was a member of the Technical Staff at AT&T Bell Laboratories, Holmdel, NJ, working in digital speech coding and enhancement techniques, digital communications systems, and radio communications. He joined the Department of Electrical Engineering, National University of Singapore, in 1983 and is currently a Senior Lecturer in the Department. In the Spring of 1985 and 1986 he was a Visiting Consultant at AT&T Bell Laboratories. His current research interests include packet voice networks, local communication systems, and combined source and channel coding techniques.