THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 7, September 1979
Printed in U.S.A.

Adaptive Aperture Coding for Speech Waveforms—I

By N. S. JAYANT and S. W. CHRISTENSEN

(Manuscript received December 29, 1978)

In aperture coding, one refrains from encoding waveform samples until the waveform crosses an appropriately wide aperture centered around the last encoded value. If the waveform is slowly varying in some sense, the above procedure can be a basis for bit rate reduction. The identification of aperture-crossing samples can be either explicit or implicit, and it is the latter case that this paper mainly addresses. We follow a finite length, converging-aperture procedure proposed recently for picture waveforms, and show that it can be used for speech coding as well if the aperture width is designed to be syllabically adaptive. We also describe, for Nyquist-sampled speech, desirable designs for aperture shape and aperture length L. The special case of $L = 1$ corresponds to ternary delta modulation with a constant encoding rate of $\log_2 3 \sim 1.6$ bits/sample. Using longer apertures (e.g., $L = 2, 3$), we show that it is possible to obtain average encoding rates as low as 1.2 bits/sample without significantly changing output speech quality. With 8- to 12-kHz sampling , the average bit rate would then be 9.6 to 14.4 kb/s. At these transmission rates, adaptive aperture coding, used in conjunction with a simple (first-order) adaptive predictor, can provide communications quality speech.

I. INTRODUCTION

The encoding technique described in this paper is intended to be a simple time-domain approach for encoding speech waveforms at transmission rates like 9.6 or 16 kb/s. The digital speech output resulting from this technique, or simple modifications thereof,[6] is expected to be of *communications* quality: less than *toll* quality, but nevertheless adequate for many applications.

The notion of aperture coding, per se, is not new. It has been considered extensively for digitizing telemetry data, with a view to

exploiting their slowly changing characteristics.[1-3] The point of this paper is that aperture coding can be useful for low-bit-rate digitizations of speech waveforms as well, provided the coding procedure is designed to be properly adaptive to the changing statistics of speech inputs. In fact, an important contribution of this paper is the specification of a rather carefully designed syllabic adaptation algorithm for aperture width.

Adaptive aperture coding is inherently a variable rate procedure, and for use with a transmission channel that expects a constant-rate output, one would need an appropriate buffer at the coder output. Typical buffer lengths and consequent encoding delays can be several tens of milliseconds. This will be of no concern when aperture coding is used for digital speech storage but, for transmission applications, the encoding delay will be an important consideration.

II. APERTURE CODING

The basic notion can be explained with reference to Fig. 1. Assume that the waveform sample at time 0 has been encoded and transmitted. The idea now is to view the immediate future of $X(t)$ through an aperture of width 2A, centered on the circle that represents the transmitted value at time 0; and to refrain from transmitting samples that lie within this aperture; the next transmission will therefore occur at time 3, after which the process continues with an updated aperture. Here, and in the next figure, open circles represent transmitted values, while solid dots denote samples deemed redundant. In reconstructing the waveform, redundant samples can be assigned amplitudes equal (for example), to that of the last transmitted sample, as shown by the dashed horizontal running through the aperture. This procedure en-

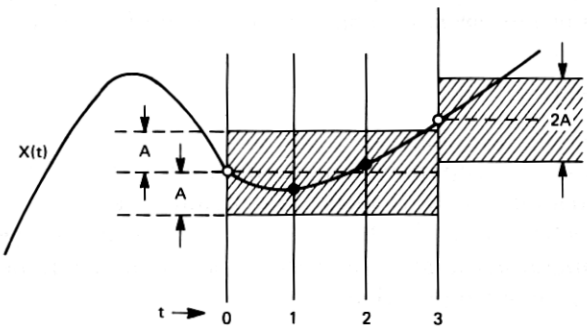

Fig. 1—Illustration of the aperture coding concept.

tails a distortion that can be referred to as aperture noise. As one increases A, aperture noise increases, but so does the proportion of samples that need not be encoded/transmitted. The tradeoff between noise and transmission probability depends on how slowly the input waveform varies, and for nonstationary inputs such as speech waveforms, the best tradeoffs are realized in schemes where one adapts A to changing input statistics. [With nonadaptive aperture schemes for Nyquist-sampled speech, a transmission probability of 1 out of 2 (or 2.5, 3, 4, 5) samples implies typical signal-to-aperture-noise ratios of about 33 (or 21, 18, 14, 11) dB, assuming that the only silences present in the speech input are naturally occurring microsilences, and not explicit pauses.]

Practical aperture schemes present two considerations which have not been introduced in Fig. 1. First, the "transmitted" samples have to be digitized somehow, so that the quality of reproduced speech will be characterized by this digitization—or quantization—noise, in addition to the aperture noise mentioned earlier. Second, the decoder at the receiving end has to know which of the input samples have been deemed redundant by the encoder, and which of them have been explicitly digitized. Most aperture coding literature[1-3] assumes explicit transmission of the above "timing" information. For example, the encoder can transmit, for each input sample, a binary number which tells the decoder whether that sample is being encoded or deemed reundant. If the probability of a nonredundant sample is p and if such a sample is further encoded using B bits, the average transmission rate is $[p \cdot B + 1]$ bits/sample, where the term 1 is due to the constant timing information bit; and the savings, relative to a zero-aperture scheme, are $[B(1 - p) - 1]$ bits/sample. This formula suggests that B has to be large enough (for a given p) so that the savings is positive in spite of the timing information. On the other hand, in low bit rate applications, values of p (that are compatible with a tolerable amount of aperture + quantization noise) may be such that the savings due to aperture coding are either insufficient or negative—unless, of course, the timing information overhead can be avoided altogether. An aperture scheme which does precisely this was described recently by Murakami, Tachibana, Fujishita, and Omura[4] in the context of picture coding, and the purpose of this paper is to describe our modification of that scheme for encoding speech waveforms with $B = 1.2$ to 1.6 bits/sample, a range of bit rates which clearly cannot afford explicit transmission of timing information.

Succeeding sections describe our findings concerning aperture characteristics that are desirable for low bit rate speech coding. These characteristics include aperture shape, aperture length (to be defined presently), and adaptation algorithms for aperture width A.

III. APERTURE CODING WITHOUT EXPLICIT TRANSMISSION OF TIMING (TIME OF NONREDUNDANT SAMPLE) INFORMATION

Consider the procedure of Fig. 2. The converging nature of the aperture is desirable, as we shall note later, but the convergence is not critical from the timing information viewpoint. As in Fig. 1, a transmitted sample at time 0 is followed by two redundant samples. The nonredundant sample $X(3)$ is encoded as follows. First, it is quantized to a level corresponding to the *previous nearest point* on the aperture characteristic ($P3$, in this case), and this value is transmitted by means of a code word dedicated to point $P3$ of the characteristic. Reception of this code word conveys two items of information to the receiver: first, that a nonredundant sample was encountered *after* $P3$, i.e., at time 3; second, that this sample has been quantized to an amplitude that is equal to that of $P3$ itself (as defined relative to the dashed horizontal running in the middle of the aperture). Once again, as in Fig. 1, the process is repeated with the transmitter (and receiver) beginning a new aperture centered on $Y(3)$, the approximation to $X(3)$.

In the above example, a (positive-sided) aperture crossing was observed at time $t = 3$. (This event will be denoted when needed as a "run" of length $R = 3$.) If the crossing was observed at time $t = 1$ on the other hand (run length $R = 1$), the input $X(1)$ would have been encoded as a value $Y(1)$, and this would have been represented by a code word (and amplitude) corresponding to $P1$ or $N1$, depending on whether the crossing was above or below the aperture center. If, on the other hand, there was no crossing even as late as $t = 3$ (run length $R > 3$), $X(3)$ would be encoded by the central "zero" level Z at the end of the aperture and a new aperture would be created, centered on $Y' = Z$.

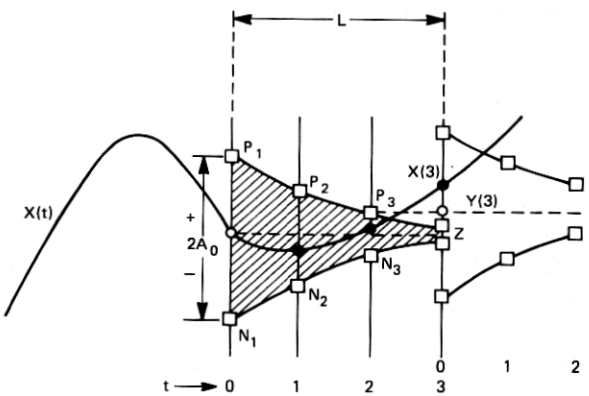

Fig. 2—Aperture coding without explicit transmission of timing information.

Table I—Aperture characteristics for L = 3 and J = 0.5

Run Length R (First time that crossing is observed)	Updating Relative to Predicted Value
1	$\pm A_0$
2	$\pm A_0/\sqrt{2}$
3	$>A_0/2$
">3" / No crossing observed	0

The use of a "zero" output level implies the use of a finite-length aperture. In fact, the aperture length can be defined as the time at which the aperture is truncated with a zero output level Z. In Fig. 2, $L = 3$, and in the particular example that has been sketched, $R = 3$ as well.

The number of "output" points on the aperture characteristic of Fig. 2 was 7 ($3P$'s + $3N$'s + $1Z$). In general, the aperture characteristics in our scheme are described by $(2L + 1)$ distinct outputs, and corresponding transmission code words.

Relative amplitudes on the characteristic are determined by the shape of the aperture. We have found that converging apertures that are appropriate for speech can be conveniently formalized into the class

$$A(t) = A_0 \cdot 2^{-J \cdot t}, \tag{1}$$

where $A(t)$ is the aperture width at time t. We have further found that a desirable range for J is $(0.5 \leq J \leq 1)$. (We have also looked at shapes described by complete convergence, $A(L) = 0$, with corresponding $(2L + 1)$-point characteristics, but we have found them to be less useful than those described by the exponential decay above.)

Table I defines the quantization characteristics of aperture coding for the illustrative case of $L = 3$ and $J = 0.5$. Notice that the output (quantized) amplitudes are defined relative to a "predicted value." In the examples of Figs. 1 and 2 we have assumed that all predicted values are equal in amplitude to the last (explicitly) transmitted amplitude, as indicated by the dashed horizontals running through the aperture areas. The situation corresponds, formally, to a first-order predictor with a coefficient a_1 equal to unity.* In general, however, one can use a speech-specific predictor $a_1 = 0.85$,[5] or a higher-order predictor (for example, $a_1 = 1.10$, $a_2 = -0.28$, $a_3 = -0.08$; see Ref. 5) to reconstruct redundant samples, and to predict nonredundant samples prior to updating, as in Table I. The coding procedures in these more general cases would be qualitatively described by Fig. 3. Further, the predictor can also be adaptive, to follow the changing spectral char-

* Nonpredictive aperture coding results if $a_1 = 0$.

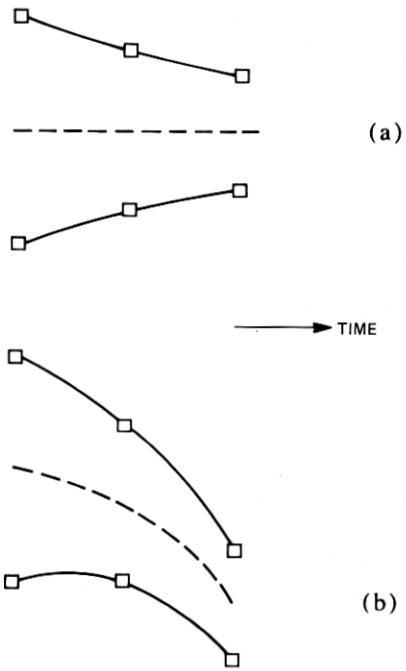

Fig. 3—Aperture coding with (a) first-order prediction: $a_1 = 1$, and (b) more general prediction.

acteristics of input speech. The adaptive predictors considered in this paper are of first order, in the interest of simplicity: the adaptive predictor coefficient is simply set equal to the one-sample-lag autocorrelation value c_1 of the speech input. The parameter c_1 is updated once for every 256-sample input block. Explicit transmission of c_1 values to a receiver will typically entail an additional information transmission of about 100 to 200 bits per second. This extra transmission can be entirely avoided in schemes where c_1 is estimated from a past history of reconstructed, rather than input, speech.[6]

IV. ADAPTIVE APERTURES

Nonadaptive and adaptive apertures are sketched in Fig. 4. The figures show the time evolution, if any, of the maximum aperture width $A_0(A(0))$. For a nonstationary signal such as speech, it is critical to have an adaptive procedure such as in Fig. 4c. The adaptations would let A_0 follow changing input statistics and provide individually tailored arrangements for encoding high-level voiced segments, low-level sounds such as fricatives, and zero-level microsilences. The results of

this paper assume that the aperture shape as described by J in (1) is fixed, and that only the width A_0 is adaptive.

We studied many adaptation algorithms, including those that can be described as instantaneous, periodic, and syllabic. The best results were obtained with syllabic adaptations as typified by the algorithm

$$A_0^{(r+1)} = G_1 \cdot A_0^{(r)} + G_2 \cdot (\text{ADAPT})^{(r)}$$

$$G_1 = 1 - \epsilon^2; \qquad \epsilon \to 0$$

$$(\text{ADAPT})^{(r)} = 1 \quad \text{if} \quad \sum_{s=1}^{4} R_{r-s} < K$$

$$= 0 \text{ otherwise.} \tag{2}$$

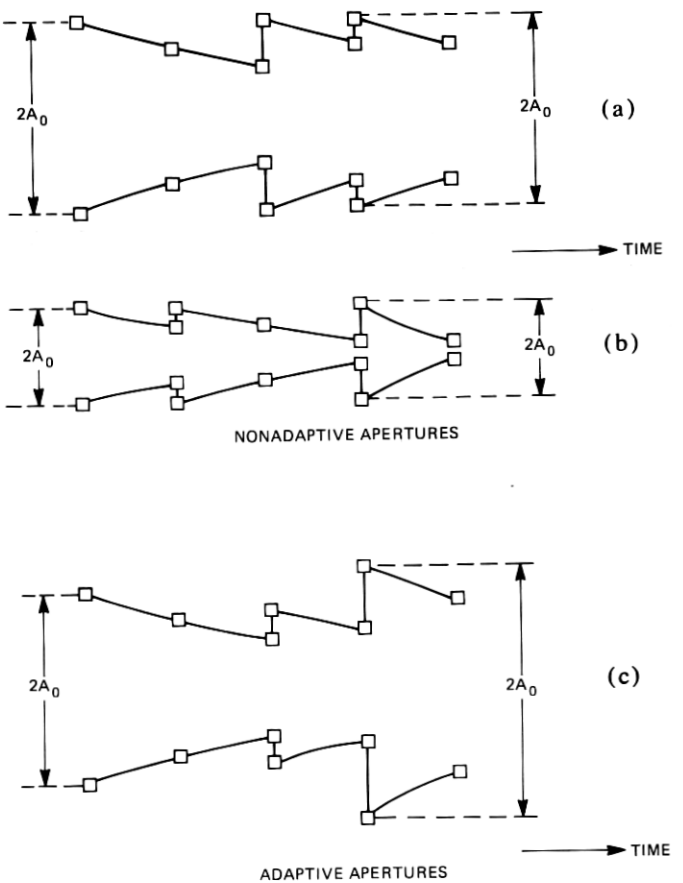

Fig. 4—Nonadaptive and adaptive apertures.

In the above algorithm, r indexes successive new apertures, and not successive input samples. In other words, going from r to $r + 1$ could mean an interval up to L input samples. The parameter R refers to the run length of redundant samples. By defining a "Z" event as a run of length $L + 1$, the parameter R is seen to have a range $(1 \leq R \leq L + 1)$. Briefly, the above adaptation logic uses a succession of four small runs as a cue for increasing A_0; in the absence of such a cue, the logic lets A_0 decrease slowly, at a rate given by G_1. Our experiments have shown that desirable values of G_1 for 8-kHz-sampled speech are between 0.95 and 0.99 (corresponding to syllabic time constants of 2 to 8 ms for aperture decreases); while good choices for the threshold K are 6, 6, and 8 for aperture lengths of $L = 1$, 2, and 3 respectively. The most interesting parameter in (2) by far was the quantity G_2 that determines the nature of aperture width increases, and we shall come back to this parameter presently.

Meanwhile, Figs. 5, 6, and 7 provide illustrative descriptions of the adaptive procedure described in (2). Figure 5 shows how the aperture width A_0 (5b) tracks input speech power (5a), Fig. 6 provides a typical histogram of A_0 samples, showing a microsilence-related concentration at very low values of A_0, and Fig. 7 compares typical aperture-crossing sequences (redundant and nonredundant samples, versus time) in nonadaptive (7b) and adaptive (7c) schemes. In the 2-state sequences of Figs. 7b and 7c, a zero state represents a redundant sample, while a nonzero state denotes an aperture crossing, or nonredundant sample.

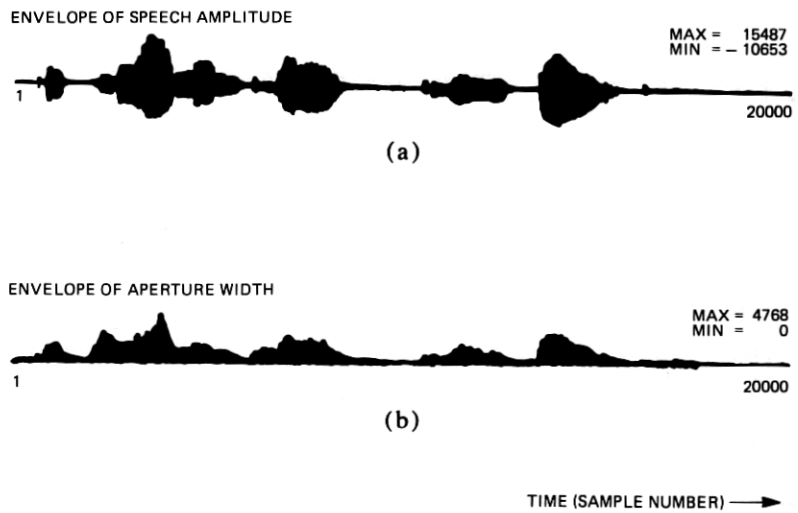

Fig. 5—Syllabically adaptive apertures: envelopes of (a) sentence-length speech (b) aperture width A_0.

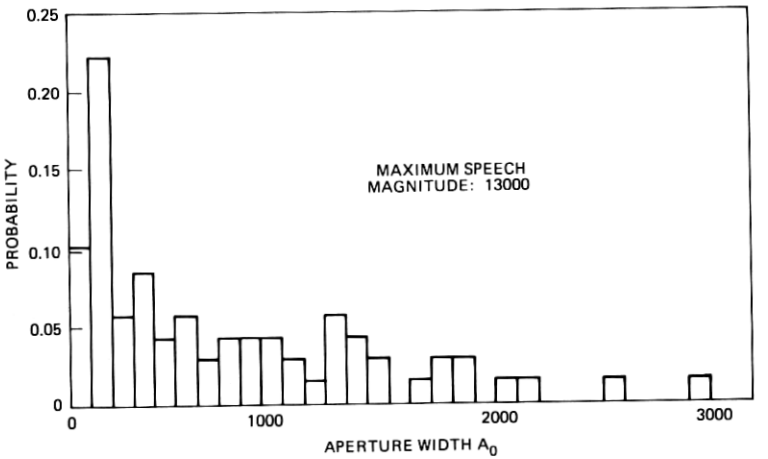

Fig. 6—Histogram of aperture width samples.

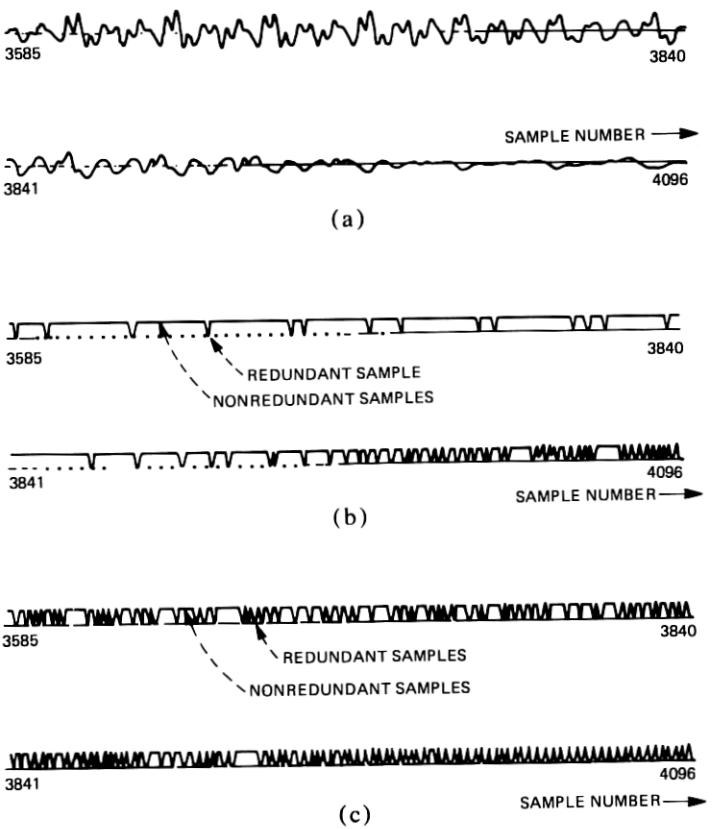

Fig. 7—Aperture crossings in (b) nonadaptive and (c) adaptive schemes, corresponding to a (a) speech waveform segment.

V. PERFORMANCE OF APERTURE CODING AS A FUNCTION OF L AND G_2

The more interesting results of our experiments (computer simulations) are summarized in Figs. 8 and 9. These results apply to a bandlimited (200 to 3200 Hz) female utterance, "THE CHAIRMAN CAST THREE VOTES," and the nonadaptive third-order predictor [$a_1 = 1.10$, $a_2 = -0.28$, $a_3 = -0.08$] mentioned earlier. The case of adaptive prediction is discussed briefly in Section VI.

5.1 Segmental signal-to-noise ratios

The objective speech quality measure used in Fig. 8 is the segmental signal-to-noise ratio SNRSEG obtained by computing the s/n ratios in 256-sample (32 ms) blocks, expressing the values in decibels, and taking the average of local decibel values over the length of the sentence-length input—a procedure which reflects low-level speech rendition better than the conventional average s/n ratio. It is significant that the maximum performance with $L = 3$ is nearly 1 dB below the peak performance with $L = 1$ and $L = 2$; and that for a given value of G_2, $L = 2$ tends to perform better than $L = 1$ (except if $G_2 < 25$). It will be seen, on the other hand, that, transmission-rate-wise, interesting values of G_2 are quite different for different values of L, and we presently reexamine the results of Fig. 8 taking the above fact into account.

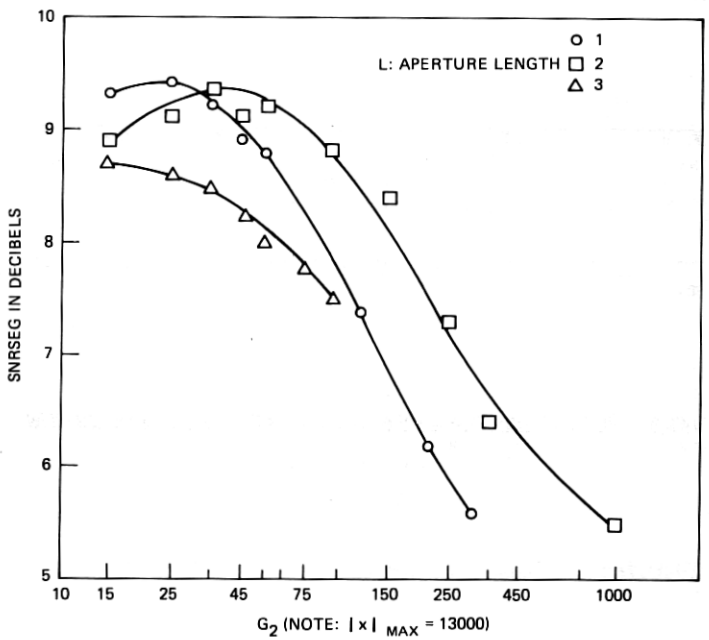

Fig. 8—Variation of speech quality SNRSEG with adaptation parameter G_2, for $L = 1$, 2, and 3 (8-kHz sampling; nonadaptive prediction).

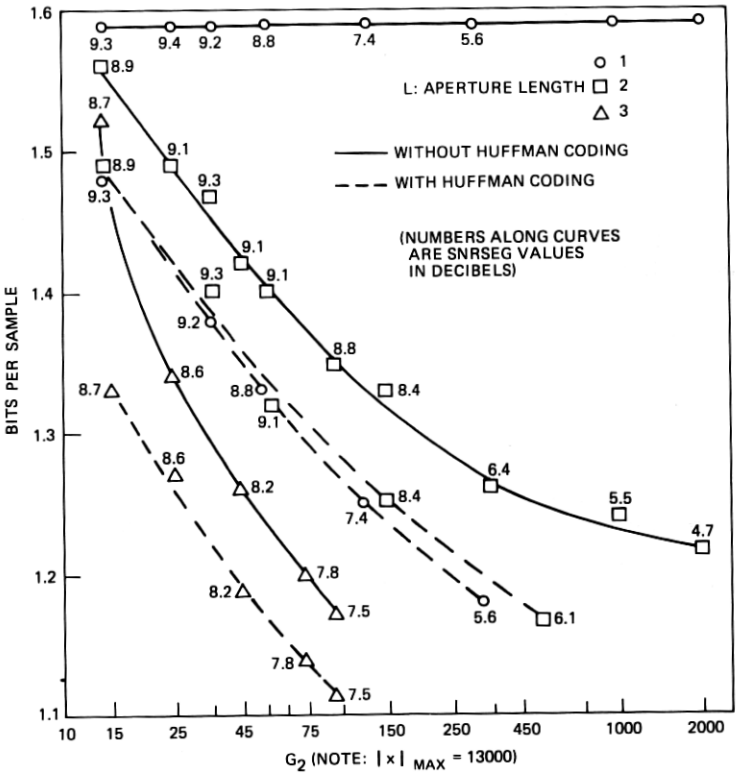

Fig. 9—Variation of transmission rate (bits/sample) with adaptation parameter G_2, for $L = 1$, 2, and 3 (8-kHz sampling; nonadaptive prediction).

Meanwhile, it should be noted that, for a given value of L, and in the neighborhood of an SNRSEG-maximizing G_2 value, increasing G_2 tends to make the output speech more granular and harsh; while decreasing G_2 tends to make the speech more low-passy and muffled. Finally, the "design" parameter in Fig. 8 is strictly the ratio of G_2 to $|X|_{max}$, the maximum input speech magnitude, rather than the absolute value of G_2.

5.2 Average transmission rates

The *average* (information) transmission rate in an aperture coding scheme is upper-bounded in the form

$$I(L) \leq p \cdot \log_2(2L + 1) \text{ bits/sample}, \tag{3}$$

where p is the probability of a nonredundant sample and $(2L + 1)$ is the number of distinct output points on the aperture characteristic.

The inequality above recognizes the fact that the $(2L + 1)$ output

code points, in general, have unequal probabilities of being used, so (for a given $p < 1$) further information compression can be achieved by assigning relatively short code words to frequent outputs and using relatively long code words for the infrequently occurring outputs. Thus, in the example of Table II, the variable-length Huffman coding results in an average bit rate of $1.34 < 1 \cdot \log_2 3 = 1.59$ bits/nonredundant sample. Note that for $L = 1$, $p = 1$ by definition because the 3-point aperture characteristic always puts out a nonredundant output corresponding to $P1$, $N1$, or Z. In fact, this special case is no more than *ternary (3-level) delta modulation*, and the output has a *constant* rate of $\log_2 3 \sim 1.6$ bits/sample. The effect of Huffman coding, however, is to make the output bit rate variable. With $L = 2$ or 3, on the other hand, the output rate is variable even without Huffman coding, because $p < 1$ in general, for these cases. Information rates for $L = 1$, 2, and 3 are shown in Fig. 9 as a function of G_2, with and without entropy (Huffman) coding in each case.

Figure 9 also includes, for convenience, the SNRSEG information from Fig. 8. For an average bit rate of 1.6 bits/sample, ternary delta modulation ($L = 1$) without Huffman coding is an obvious choice: there is no motivation for aperture coding and the attendant variability in the encoder output rate. For average bit rates of about 1.4 bits/sample, one has the choice: $L = 1$ with Huffman coding or $L = 2$ without Huffman coding. It is apparent that, for the greatest reductions of information rate (say, $I(L) = 1.2$ bits/sample), one needs to employ nontrivial ($L > 1$) aperture coding, an observation that is also suggested by the literature on adaptive asynchronous delta modulation.[7] In our scheme, the justification for $L = 3$ comes directly from the fact that values of G_2 that realize 1.2 bits/sample encoding are far too suboptimal (SNRSEG-wise) in the cases of $L = 1$ and $L = 2$ (see Fig. 8).

VI. ADAPTIVE PREDICTION

We have studied the performance of an adaptive aperture coding scheme where the waveform predictor is also adaptive. In the interest of simplicity, we have confined our studies to the case of first-order prediction. In this case, the adaptive prediction procedure is simply to compute the adjacent sample correlation c_1 of input speech samples X

Table II—Huffman coding example (L = 1, G_2 = 45)

Sign	Run Length	Probability	Code Word
+	1	0.17	0 0
−	1	0.17	0 1
	>1	0.66	1
Transmission Rate	$= 0.17 \cdot (2) + 0.17 \cdot (2) + 0.66 \cdot (1)$ $= 1.34$ bits/sample		

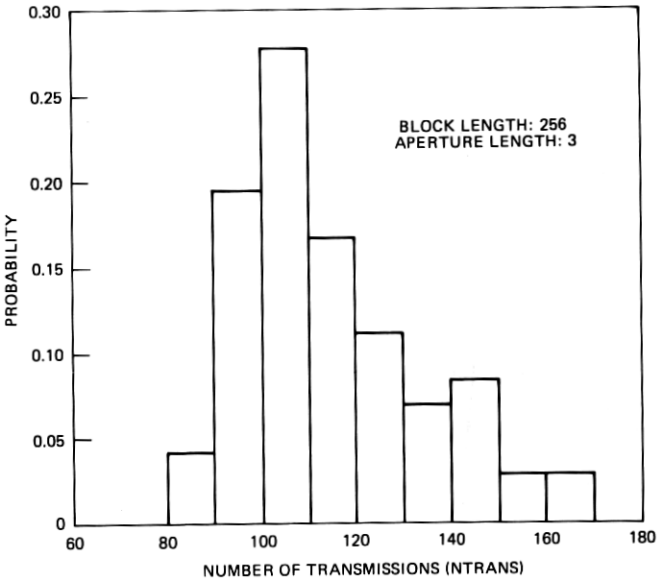

Fig. 10—Histogram of the number of transmissions in a 256-sample block ($L = 3$).

(typically, once for each input block of 256 samples), and to set the predictor coefficient a_1 equal to c_1:

$$a_1 = c_1 = \frac{E[x_r x_{r-1}]}{E[x_r^2]}; \qquad E(\cdot): \text{expected value.} \tag{4}$$

The use of adaptive prediction did not increase the SNRSEG values of Fig. 9 drastically, but perceptual improvements in coded speech quality were quite significant. The resulting speech output with 1.2 to 1.6 bits/sample aperture coding has *communications* quality: the degradation is obvious in a direct comparison with the input speech, but the quality should nevertheless be adequate for many communication purposes. The output speech quality also varies with input speech: with certain types of input, the output speech is highly *intelligible* even with nonadaptive prediction and 8-kHz sampling. The speech *quality*, however, improves significantly with adaptive prediction and faster sampling (say, 12kHz), and with adaptive low-pass filtering of the output.[6] Finally, in informal comparisons with adaptive delta modulation (ADM) at a given bit rate, adaptive aperture coding is clearly better, as expected.

VII. VARIABILITY OF BIT RATE IN THE APERTURE CODER OUTPUT

Aperture coding schemes, with the exception of the special case of ternary delta modulation without Huffman coding, generate variable-rate outputs. For example, Fig. 10 shows a histogram of sample values

of NTRANS, the number of transmissions per 256-sample block, in a scheme with $L = 3$. Note that the nonredundant sample probability p varies in the range $0.31 \leq p \leq 0.66$.

If a variable-rate procedure is used to decrease the average bit rate, one needs the additional provision of a bit buffer to be able to deliver bits at a constant rate into a channel (that accepts them in that format). The necessary length of such a buffer can be equated to the peak-to-peak variation of the quantity

$$\sum_{u=1}^{N} B_u - \bar{B}.N, \tag{5}$$

where B_u is the number of bits used to encode speech sample $u (B \geq 0)$, N is the total number of speech samples in a (statistically long enough) test input, and $\bar{B}$ is the average bit rate (bits/sample) needed to transmit that test input.

Using the sentence-length utterance mentioned earlier, we evaluated the peak-to-peak excursion of (5) for three cases: (*i*) $L = 1$ plus Huffman coding, $G_2 = 35$; (*ii*) $L = 2$ without Huffman coding, $G_2 = 35$; and (*iii*) $L = 3$ with Huffman coding, $G_2 = 45$. Cases (*i*) and (*ii*) correspond to $\bar{B} = 1.4$, and case (*iii*) is $\bar{B} = 1.2$. Respective buffer requirements were approximately 600, 400, and 800 bits. Respective encoding delays (for 8-kHz sampled speech and appropriate $\bar{B}$ values) are approximately 50, 35, and 80 ms.

For speech transmission applications, the above delays are significant if not prohibitive. Furthermore, in practical designs of aperture coding, one should specify a maximum buffer length, and they should include an automatic procedure[8] for increasing or decreasing the local average rate $\bar{B}$, depending on current buffer status as given by (5). Clearly, the parameter G_2 would be a natural means for controlling local values of $\bar{B}$.

In multiplex-speech situations, active (high output bit-rate) and inactive (low output bit-rate) segments get more intermixed in time than with a single speech channel. Consequently, buffering problems are expected to be less severe with multiplex-speech inputs. In fact, there is at least one "digital TASI" application, SPEC (Speech Predictive Encoded Communications), which indeed employs a simple form of aperture coding.[9]

The most straightforward application of aperture coding will perhaps be in the context of speech storage. In storage applications, encoding delays are less objectionable than in transmission, and buffer overflow problems, if any, need not be combatted in real time.

REFERENCES

1. C. M. Kortman, "Redundancy Reduction—A Practical Method of Data Compression," Proc. IEEE, *55*, No. 3 (March 1967), pp. 253–263.

2. L. Ehrman, "Analysis of Some Redundancy Removal Bandwidth Compression Techniques," Proc. IEEE, *55*, No. 3 (March 1967), pp. 278–287.
3. L. D. Davisson and R. M. Gray, ed., *Data Compression*, Halsted Press, 1976.
4. K. Murakami, K. Tachibana, H. Fujishita, and K. Omura, "Variable Sampling Rate Coder," *Technol. Report*, Vol. 26, Univ. of Osaka, Japan, October 1976, pp. 499–505.
5. N. S. Jayant, "Pitch-Adaptive DPCM Coding of Speech with Two-Bit Quantization and Fixed Prediction," B.S.T.J., *56*, No. 3 (March 1977), pp. 439–454.
6. N. S. Jayant and S. A. Christensen, "Adaptive Aperture Coding of Speech Waveforms—II," unpublished work.
7. T. A. Hawkes and P. A. Simonpieri, "Signal Coding Using Asynchronous Delta Modulation," IEEE Trans. on Comm., *COM*-22, No. 5 (May 1974), pp. 729–731.
8. J. J. Dubnowski and R. E. Crochiere, "Variable Rate Coding of Speech," B.S.T.J., *58*, No. 2 (February 1979), pp. 577–600.
9. S. J. Campanella and J. A. Sciulli, "Speech Predictive Encoded Communications," Second International Conference on Digital Satellite Communications, paper E4, Paris, November 1972.