THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Simultaneous Transmission of Speech and Data Using Code-Breaking Techniques

By R. STEELE and D. VITELLO

(Manuscript received April 19, 1981)

A system whereby speech is used as a data carrier is proposed. The speech, sampled at 8 kHz, is divided into blocks of N samples, and provided the correlation coefficient and mean square value of the samples exceed system thresholds, data is allowed to be transmitted. If the data is a logical 0, the samples are sent without modification; however, if a logical 1 is present, frequency inversion scrambling of the samples occurs. The receiver performs the inverse process to recover both the speech and data. Data rates of 700 b/s were achieved without data errors or speech distortion via an ideal channel. The effects of additive background and channel noise were investigated, and the system was shown to operate at 126 b/s with no data errors when the additive noise was as high as 10 dB below the mean square value of the speech signal.

I. INTRODUCTION

There are numerous schemes[1,2] for analog scrambling of speech signals, but they all require a scrambling key. For example, we may sample the speech at a rate in excess of its Nyquist rate, parcel the samples into blocks, and rearrange the blocks prior to transmission. This rearrangement of the blocks breaks up the rhythm in the speech making it difficult for an eavesdropper to comprehend the conversation. The shuffling of the block positions is done under the auspices of the scrambling-key, and provided the receiver knows this key and, hence, the descrambling key, the blocks of speech can be correctly repositioned and made intelligible to the desired recipient.

It is not our purpose to describe the numerous scrambling techniques, but rather to suggest a method whereby speech and data can be transmitted simultaneously over the channel by using scrambling

strategies. The principle is very simple. The scrambling key becomes the data to be transmitted. The receiver adopts the role of code-breaker. Every time the receiver correctly guesses the key and breaks the code, it recovers both the speech and the data. For the scheme to have any significance, the receiver must break the code successfully at nearly every attempt. Therefore, we must select scrambling keys which are easy to break, and this means that we are not aiming for speech privacy (although a degree of privacy may be achieved as a by-product). The scrambling process is, therefore, a catalyst which enables the data to be transmitted.

At first sight, it might appear that we are getting something for nothing. With care we can arrange for the data to be transmitted at negligible error rate, the speech faithfully recovered, and a small bandwidth expansion of the transmitted signal compared to the original speech. These rewards are derived from the inherent redundancy in the speech signal. Indeed, we emphasize that the method will work with any signal that has correlative features, such as speech, television, facsimile, and analog-plant control signals, like pressure and temperature variations, etc.

II. SIMULTANEOUS SPEECH AND DATA TRANSMISSION USING FREQUENCY INVERSION SCRAMBLING

As a demonstration of the concept, we describe the transmission of data using the simplest of scrambling methods, frequency inversion. In this method, speech, band-limited to 3.4 kHz, is sampled at 8 kHz and N samples are processed at a time. Let us represent these samples as

$$S_1 = x_0, x_1, x_2, \cdots, x_{N-1}. \tag{1}$$

To invert the frequency components associated with these N samples, all we need to do is to alter the polarity of every other sample,[3,4] namely,

$$S_2 = x_0, -x_1, x_2, -x_3, \cdots, -x_{N-1} \tag{2}$$

N even.

In frequency-inversion scrambling (FIS), sequence S_2 would always be transmitted, but in our scheme, sequence S_2 is used when we decide to transmit data and, further, the data is a logical 1. Observe only one bit per N speech samples is transmitted.

To minimize the number of bits received in error, we proceed as follows. The calculation

$$\rho = \frac{\sum_{i=0}^{N-2} x_i x_{i+1}}{\sum_{i=0}^{N-1} x_i^2} \tag{3}$$

is made on the original speech sequence S_1 and called here the correlation coefficient, and the mean square value

$$\sigma_x^2 = \frac{1}{N} \sum_{i=0}^{N-1} x_i^2 \tag{4}$$

in the block of speech samples is also found. Notice that the correlation coefficient ρ_s of the scrambled sequence S_2 is $-\rho$. Figure 1 shows the block diagram of the system. Mean square value σ_x^2 and correlation coefficient ρ are compared with system threshold parameters T and K in comparators COMP 1 and COMP 2, respectively. Parameters T and K may be selected such that $\sigma_x^2 > T$ and $\rho > K$ generally implies the absence of unvoiced speech and silence, assuming there is no additive background noise. This strategy aids in reducing the number of received bit errors when transmitting through noisy channels. Later we will give details of how T and K are selected.

Data is only transmitted when the Boolean equation

$$y = C_1 C_2 \tag{5}$$

is a logical 1, where

$$C_1 = \begin{cases} \text{logical 1 } \textit{if } \sigma_x^2 \geq T \\ \text{logical 0 } \textit{if } \sigma_x^2 < T \end{cases}$$

and

$$C_2 = \begin{cases} \text{logical 1 } \textit{if } \rho \geq K \\ \text{logical 0 } \textit{if } \rho < K \end{cases}.$$

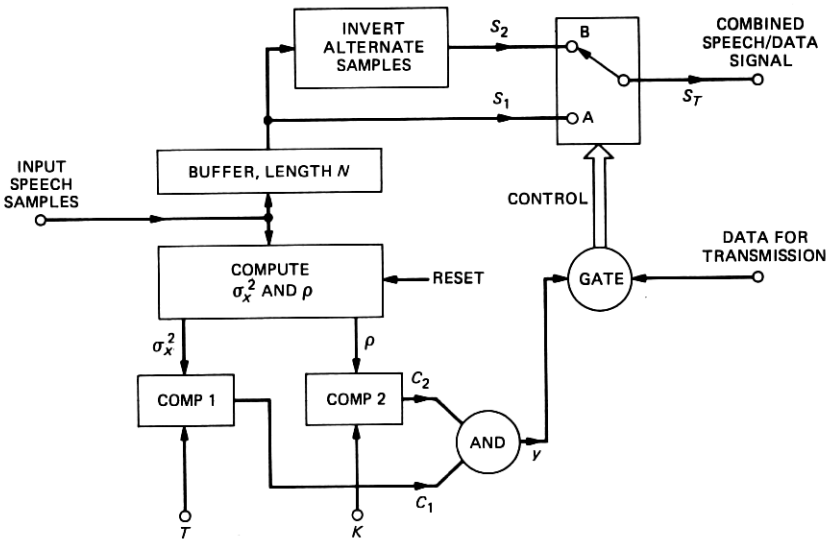

Fig. 1—Block diagram of the SSDT/FIS system at the transmitting end for the simultaneous transmission of speech and data.

The data sequence is allowed to select S_1 or S_2 if eq. (5) is satisfied. Thus, if $y = 1$, the switch in Fig. 1 is set in position A or B if the data is logical 0 or 1, respectively, i.e., a sequence S_T is generated according to

$$S_T = \begin{cases} S_1, \text{ data} = \text{logical } 0 \\ S_2, \text{ data} = \text{logical } 1 \end{cases}. \tag{6}$$

Whenever $y = 0$, $S_T = S_1$, the unscrambled speech. The sequence S_T is appropriately filtered and transmitted as the combined speech and data signal.

To illustrate the effect of the imposition of data on the speech signal, we show the waveforms in Fig. 2. In (a) and (b) of Figure 2 an arbitrary segment of speech and the corresponding transmitted signal containing data for 120 blocks are shown, respectively. The envelope of the signal is barely changed, and blocks conveying zeros are not scrambled. Hence, the transmitted signal is perceived as a distorted version of the input speech—intelligible but tiresome to a listener. A smaller segment of the original speech signal, and the resulting transmitted signal for the logical values of the data shown, are displayed, respectively, in (c) and (d) of Fig. 2. There are now only 24 blocks, and the frequency inversions are apparent when the data is a logical 1.

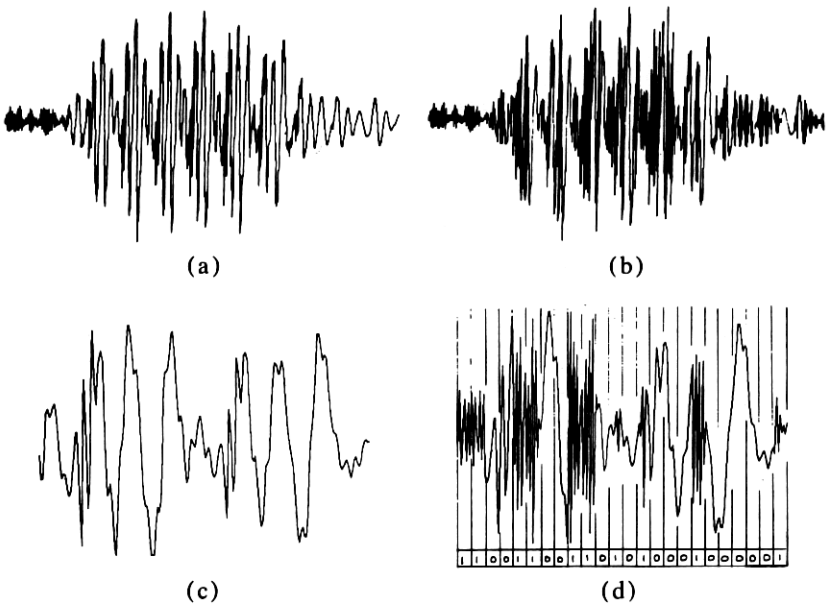

Fig. 2—Arbitrary segments of speech are shown in (a) and (c), and the corresponding transmitted signals are displayed in (b) and (d), respectively, $N = 8$, $T = 0$, $K = 0.6$. The logical values of the data signal are shown for the transmitted signal (d).

The signal emerging from the transmission channel is sampled at 8 kHz to give $\hat{S}_T$, where a caret ($\hat{}$) above the symbol signifies its presence at the receiver. In the absence of channel impairments $\hat{S}_T = S_T$, the power $\hat{\sigma}_x^2$ and correlation coefficient $\hat{\rho}$ of the sequence $\hat{S}_T$ in the block of N samples is computed according to eqs. (3) and (4). The operations associated with eq. (5) are implemented, and the following processes are performed until a decision is reached.

(*i*) If $\hat{y}$ is a logical 1, data is assumed to be transmitted of value logical 0, and $\hat{S}_T = \hat{S}_1$ is the recovered speech sequence.

(*ii*) If $\hat{y}$ is a logical 0, data may or may not be present. To determine whether data is present, every other sample in $\hat{S}_T$ is inverted and the scrambled correlation coefficient $\hat{\rho}_s$ is computed. Then,

(a) if $\hat{y}$ remains a logical 0, it is decided that no data was sent. The recovered speech sequence is, therefore, the original received sequence $\hat{S}_T$.

(b) if $\hat{y}$ becomes a logical 1, it is decided that data is present of value logical 1, and the recovered speech sequence is the scrambled $\hat{S}_T$ sequence.

Observe that if the conditions are not correct for the conveyance of data, or if a logical 0 is transmitted, the speech is dispatched without being scrambled. Only when a logical 1 is transmitted is scrambling performed, and this is done twice, once at the transmitter and once at the receiver. Should a data error occur, the speech at the output of the receiver may be erroneously scrambled. The resulting error samples in the block of length N have a rate of 4 kHz, and magnitudes double that of the original speech samples.

III. PERFORMANCE PARAMETERS FOR DATA TRANSMISSION

From a data transmission point of view we are interested in the transmitted bit rate (TBR) and the total bit error rate (TBER). Data will only be transmitted when y of eq. (5) is a logical 1, and the efficiency η of the system to transmit data is given by

$$\eta = \frac{\text{actual data rate}}{\text{possible data rate}} \tag{7}$$

from which

$$\text{TBR} = \frac{\eta f_s}{N}, \tag{8}$$

where f_s is the sampling rate of the speech signal. Error bits are those bits generated incorrectly at the receiver, and the number of bit errors per second is the TBER. Let the measure of the deficiency of the system that results in erroneous data at the output of the receiver be known

as the data transmission deficiency,

$$\lambda = \frac{\text{data error rate}}{\text{possible data rate}}. \tag{9}$$

Then

$$\text{TBER} = \frac{\lambda f_s}{N} \tag{10}$$

or

$$\text{TBER} = \text{BER} + \text{FBR}, \tag{11}$$

where BER is the conventional bit-error rate that relates to those bits transmitted that were erroneously received. The term FBR is the false-bit rate that is associated with the generation of bits at the receiver when none were actually transmitted, and the declaration at the receiver that no bits were transmitted when they really were. Representing the states when the transmitter does not transmit data, and when the receiver deems that no data was transmitted, by the symbol -1, and using the logical data symbols of 1 and 0, we are able to construct Table I, which shows all the possible data-error conditions.

Let us consider the case of no additive noise to the speech input signal, and an ideal channel. In this case, the false bits are always a logical 1 and occur when no data (-1) was transmitted. This is state A in Table I. These errors occur when the power in the block is above the threshold, $\sigma_x^2 \geq T$, and the correlation ρ is below its threshold, $\rho < K$, prohibiting transmission of data. Now K is a positive number, and the bit error will occur if ρ is negative having a magnitude K_1, say, that is greater than K. At the receiver, $\hat{S}_T = \hat{S}_1$, $\hat{y}$ = logical 0 and, hence, the received sequence is scrambled. Because the correlation

Table I—Data error table and output speech status

State	Data Status		Recovered Speech Status at RX
	TX	RX	
A	−1	1	I
B	−1	0	C
C	1	0	I
D	1	−1	I
E	0	1	I
F	0	−1	C

Note: Logical states of the data are represented by 1 and 0. When no data is sent, or no data received, −1 is used. When the output speech at the receiver for the block of N samples is correct, the symbol C is used; when it is scrambled, i.e., frequency inverted, I is used.

coefficient of the scrambled sequence is $\hat{\rho}_s = -\hat{\rho} = +K_1$, and $K_1 > K$, $\hat{y}$ is now a logical 1, and data is deemed to be present having a value logical 1. Thus, the probability of a false bit being generated is very low, being the joint probability that $\sigma_x^2 \geq T$ and $\rho < -K$.

When the speech signal is in a noisy environment, the symbols $x_{(\cdot)}$ representing speech in eqs. (1) to (4) are replaced by $x'_{(\cdot)} = x_{(\cdot)} + n_{(\cdot)}$, where $n_{(\cdot)}$ is the noise component and the $'$ above the symbols means noise contamination. The effect of the noise is to increase $\sigma_{x'}^2$ and decrease ρ', and as both $\sigma_{x'}^2$ and ρ' must exceed their thresholds [see eq. (5)] for data to be transmitted, the TBR decreases. Provided the channel is ideal, the TBER will depend on the correlative properties of the received speech, and the only source of errors derives from state A, i.e., TBER = FBR.

When clean speech is used and the channel is noisy, the TBR is unaffected. However, the TBER increases with channel noise power because the noise decorrelates the received signal, causing the receiver to sometimes erroneously presume that no data was transmitted. Thus, states D and F apply for this condition, and as the existence of other states occurs with a much lower probability, the received bit rate is approximately the difference between TBR and FBR.

Dispersive channels alter both the power and correlation of the recovered signal. The most common state is D which occurs when $\hat{\rho} < |K|$. State C occurs when the scrambled speech arrives with a correlation $\hat{\rho} \geq K$, causing 1 to be interpreted as a 0. State F occurs when $\hat{\rho} < |K|$, or $\hat{\sigma}_x^2 < T$, or when both $\hat{\rho} < |K|$ and $\hat{\sigma}_x^2 < T$. The other states were found to rarely happen.

IV. DATA TRANSMISSION PERFORMANCE

The simultaneous speech and data transmission using frequency inversion scrambling, SSDT/FIS, described here, was investigated using the sentences: "Live wires should be kept covered," and "To reach the end he needs much courage"—spoken by a male and female, respectively. The speech signal was sampled at 8 kHz to yield 38,912 samples, a number sufficiently large to give a good indication of the system's performance. The amplitude of the speech samples was confined to the range extending from −6000 to +6000 arbitrary units, and the mean square value of the samples averaged over both sentences was $MS_x = 1.09 \times 10^6$ or 60.4 dB relative to a mean square value of unity. The time waveforms for these two sentences and an expanded version of the magnitude of these speech samples to give the time variation of the low-level sounds are shown in Fig. 3.

Our objectives were to determine how to select K, T, and N for high TBR and low or negligible TBER, and to study how the performance

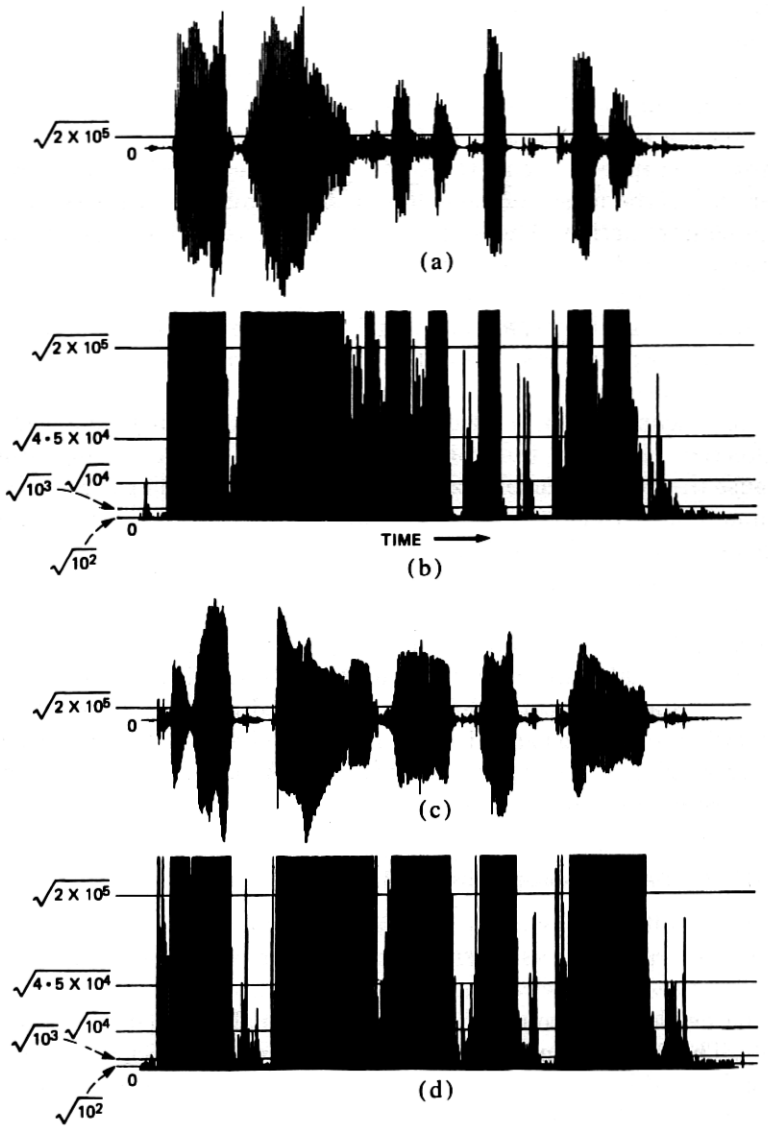

Fig. 3—Time waveforms for "Live wires should be kept covered," and, "To reach the end he needs much courage," are shown in (a) and (c). The corresponding positive amplitudes of the waveforms for the low-level sounds (high amplitudes truncated), together with various values of $\sqrt{T}$, are displayed in (b) and (d), respectively.

deteriorated in the presence of additive noise on the input speech and on the transmitted SSDT/FIS signal. We assumed that block synchronization between transmitter and receiver was correctly maintained at all times.

4.1 Selection of K

The two sentences were processed sequentially. The speech samples were divided into blocks of N samples, where N could be either 8, 16, 32, 64, 128, or 256, resulting in 4864, 2432, 1216, 608, 304, and 152 blocks of samples, respectively. For each value of N the probability density function (PDF) was computed for the correlation coefficient ρ, and plotted in Fig. 4. The PDFs were found to have similar shapes for $N =$ 16 to 256, although the shape marginally altered for $N = 8$. For smaller values of N, there is a translation in the position of the PDF to lower values of ρ. This arises because of the definition of ρ given by eq. (3). The maximum possible value of ρ for $N = 4$, 3, and 2 is 0.809, 0.707, and 0.5, respectively. We will subsequently show that $N = 4$ is the smallest block size of interest in this transmission system; therefore, we do not display PDFs in Fig. 4 for $N < 4$.

In the SSDT/FIS system, with the threshold T set to zero, the signal used to transmit data is the original speech signal, for which the curves in Fig. 4 apply. However, if $T > 0$, more blocks of speech are rejected for the conveyance of data. Therefore, TBR decreases, and the resulting blocks available for data transmission have PDFs for the correlation coefficient that are different from those in Fig. 4. At this stage, we will confine the discussion to the case of $T = 0$, i.e., where TBR has its highest values, and the curves in Fig. 4 are relevant.

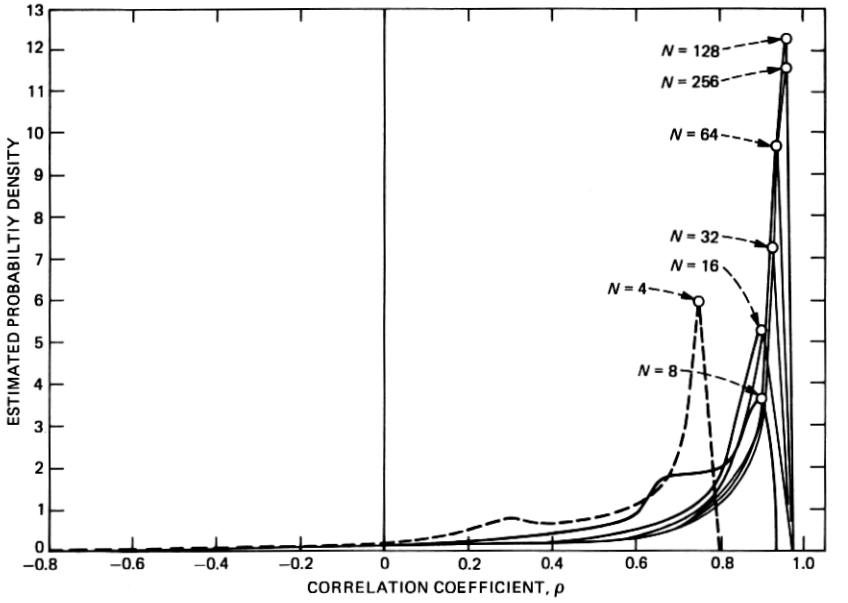

Fig. 4—Probability density function for the correlation coefficient ρ for different values of N.

Returning to these curves, we draw attention to their most negative correlation coefficient, ρ_{min}, values, as they can have a significant effect on the number of bit errors. The variation of ρ_{min}, and the maximum correlation coefficient ρ_{max}, as a function of N is displayed in Fig. 5. We recall from our discussion in Section III, that if $\rho < K$, $\sigma_x^2 > T$, no data is transmitted. Assuming an ideal channel, and given that $\hat{\rho} = \rho < -K$, the system is fooled into believing a logical 1 was transmitted and a bit error occurs. Clearly, if K is selected such that $\rho < -K$ does not exist, then no bit errors are possible over an ideal channel. To avoid bit errors we arrange for

$$K > |\rho_{min}|, \tag{12}$$

and the choice of K to avoid bit errors as a function of N must, therefore, be below the curve $|\rho_{min}|$, e.g., for $N = 64$, $K > 0.43$. For $N < 16$, ρ_{min}, and ρ_{max} both decrease with decreasing N, and for $N = 4$ we have the interesting situation that $|\rho_{min}| = \rho_{max}$, which means that if Inequality eq. (12) is satisfied no data will be transmitted as $\rho > K$ cannot exist. The value, $N = 4$, therefore, marks the lower limit of the block size for combined speech and data transmission over an ideal channel without the occurrence of bit errors.

Reducing K from ρ_{max} increases the number of speech blocks that can be considered for the conveyance of binary data, but if $K \leq |\rho_{min}|$, bit errors ensue. Thus, in order to transmit the greatest amount of

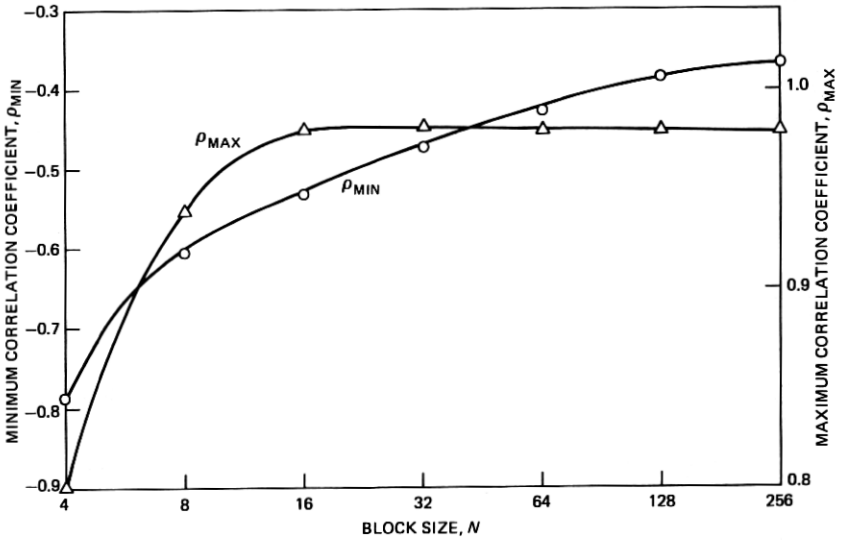

Fig. 5—Variation of maximum (ρ_{max}) and minimum (ρ_{min}) correlation coefficient values for different block sizes (N).

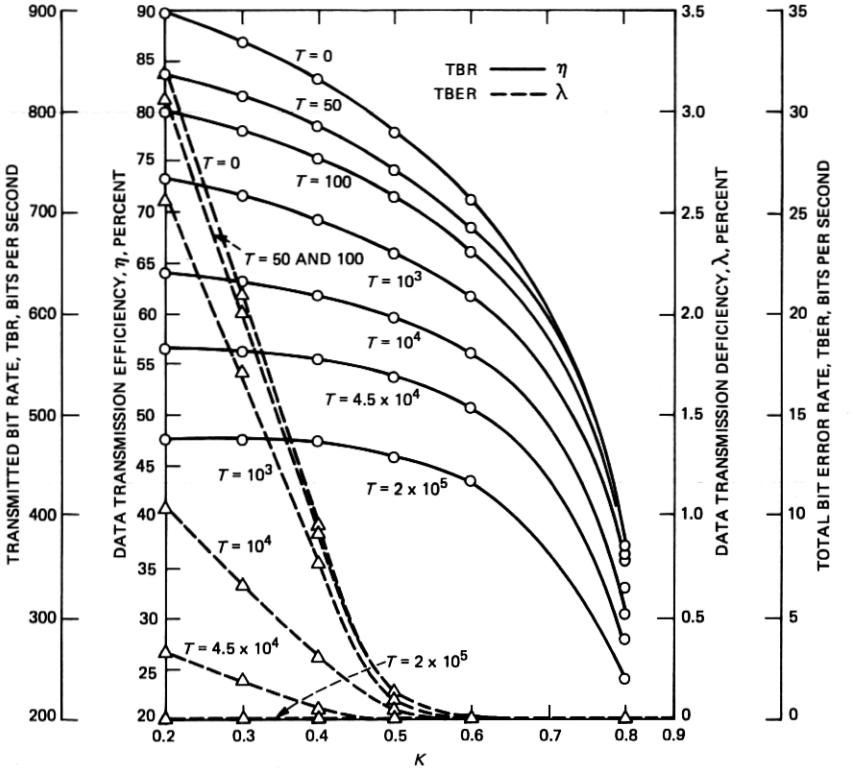

Fig. 6—Variation of data transmission efficiency (η) and data transmission deficiency (λ), as a function of K and T for $N = 8$.

data without errors over an ideal channel, K is bounded by

$$|\rho_{\min}| < K < \rho_{\max} . \tag{13}$$

However, the negative tails of the PDFs are long and of low amplitude and, hence, $K < |\rho_{\min}|$ can be used provided the penalty of a low TBER can be tolerated.

4.2 Ideal channel

The effects of parameters K and T on the data transmission efficiency η, the TBR, the data transmission deficiency λ, and the total TBER, for block sizes of 8 and 32, is shown in Figs. 6 and 7, respectively. These two block sizes were selected because $N = 8$ provides the highest data transmission rate in the absence of false bits, and $N = 32$ has fewer false bits than does $N = 8$ at low values of K, while having a relatively high TBR. Because the shape of the curves in Figs. 6 and 7 are similar, we refrain from showing curves for other values of N.

The curve for $T = 0$ is of interest as it provides the highest values of

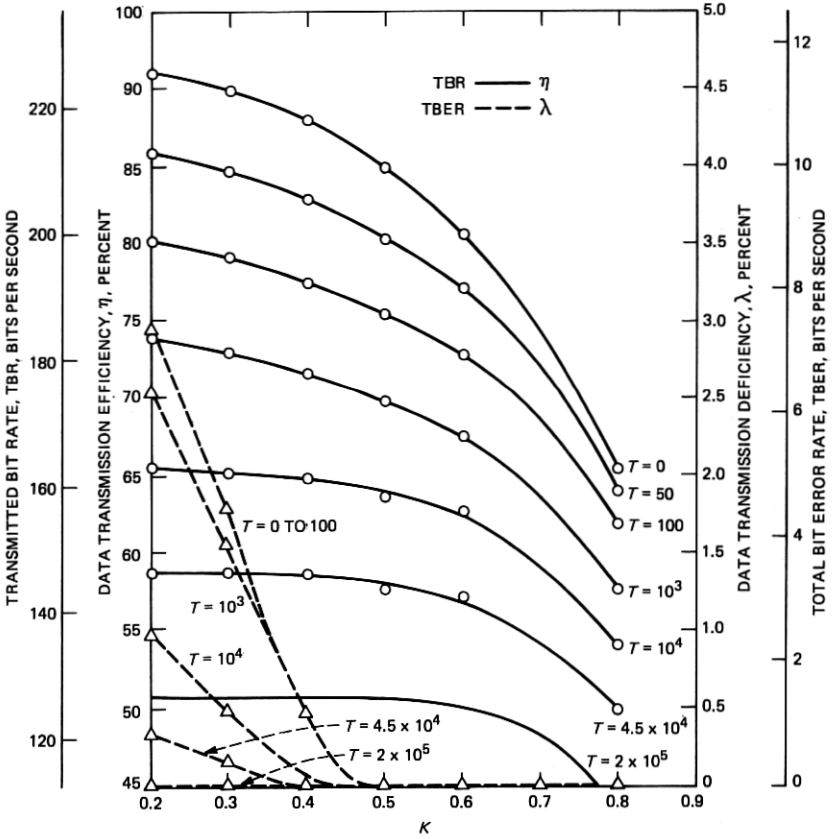

Fig. 7—Variation of data transmission efficiency (η) and data transmission deficiency (λ), as a function of K and T for $N = 32$.

η, and relates to the discussion in Section 4.1. If k is increased beyond 0.8 the curve falls rapidly, becoming zero for $K \geq \rho_{max}$. As K is reduced below 0.2, η climbs towards 100 percent, but λ becomes excessive. Consider the operating condition: $T = 0$ and $K = |\rho_{min}|$. For $N = 8$, $K = 0.6$, $\eta = 71$ percent, $\lambda = 0$, yielding a TBR of 710 b/s and a TBER of zero. By reducing K to 0.2 while maintaining $T = 0$, η is increased to 90 percent, or to a TBR of 898 b/s. However, λ is now 3.35 percent, giving a TBER = FBR of 33.5 b/s. These FBRs arise because K is below $|\rho_{min}|$. The system can operate with low values of K, 0.2 say, provided T is increased. By raising the value of T, blocks which occur during silence and unvoiced periods are not considered for data transmission. For still higher values of T, blocks existing during silence, unvoiced, and low amplitude voiced sounds, are rejected for the conveyance of data. The values of T used in our experiments (other than $T = 0$),

namely 50, 100, 10^3, 10^4, 4.5×10^4, 2×10^5; correspond to power levels that are 43.4, 40.4, 30.4, 20.3, 13.8, and 7.36 dB, respectively, below the mean square value MS_x of the combined speech signals. The $\sqrt{T}$ thresholds, except $\sqrt{50}$, are shown in Fig. 3(b) and (d), and for reference $(2 \times 10^5)^{1/2}$ is shown in Fig. 3(a) and (c).

The effect of using non-zero values of T is a modification of the shape of the PDF of the correlation coefficient ρ—specifically, the truncation of its long negative tail. Consequently, $\rho_{\min}$, a negative value in Fig. 3, is made significantly more positive. For example, when $N = 8$, $\rho_{\min} = -0.6$, -0.56 and -0.07 for $T = 0$, 10^3 and 2×10^5, respectively. When $T = 2 \times 10^5$, no false bits occur, irrespective of K, as shown in Fig. 6, but η decreases significantly to 48 percent, giving a TBR of 480 b/s. Clearly, for the ideal channel, there is no advantage in making T anything other than zero and $K = 0.6$. We will find that in the presence of channel noise T must have a high value if λ is to be contained.

Although increasing N generally produces higher values of η, as can be seen in Fig. 7, the larger block size results in a significant reduction in TBR. The effect of N on λ is seen to be small; therefore, we recommend the use of $N = 8$.

The effect of block size N on TBR and TBER for different values of T is shown in Fig. 8, where $K = 0.5$. The data transmission efficiency is approximately independent of N for a given T, and consequently TBR is inversely proportional to N. [See eq. (8)]. False-bit errors occur for $N = 16$ and 8 as $\rho_{\min} < -0.5$, unless T is increased to $\simeq 4.5 \times 10^4$. By using this high value of T, TBR is seen to fall from 530 b/s to 19.5 b/s as N is increased from $N = 8$ to 256, the TBER being maintained at zero. Clearly, from a data transmission point of view, high values of N are undesirable, although they do increase the listening fatigue of an eavesdropper, as described in Section 5.2.

The small block size of $N = 4$ that spans a duration of 0.5 ms can be used to transmit data without error, provided a large value of T is used to remove the long tail in the correlation coefficient PDF of Fig. 4. We found that if $T = 2 \times 10^5$, $\lambda = 0$, and $\eta = 31.6$ percent. This represents a TBR of 632 b/s and a TBER of zero.

4.2.1 *Effect of background noise*

To simulate a noisy environment, we added a random noise sequence having a power σ_{ni}^2 to the speech sequence. The effect of this background noise power on the data transmission efficiency η and TBR for different values of threshold T is shown in Figs. 9 and 10 for $N = 8$, $K = 0.6$, and $N = 32$, $K = 0.5$, respectively. An additive power level of $\sigma_{ni}^2 = 10^k$, $k = 0, 1, 2, \cdots$, corresponds to a power level of 60.4-10 k, dB, below the mean square value MS_x of the speech signal. As expected, the highest value of η occurs when $T = 0$, as the only criterion applied

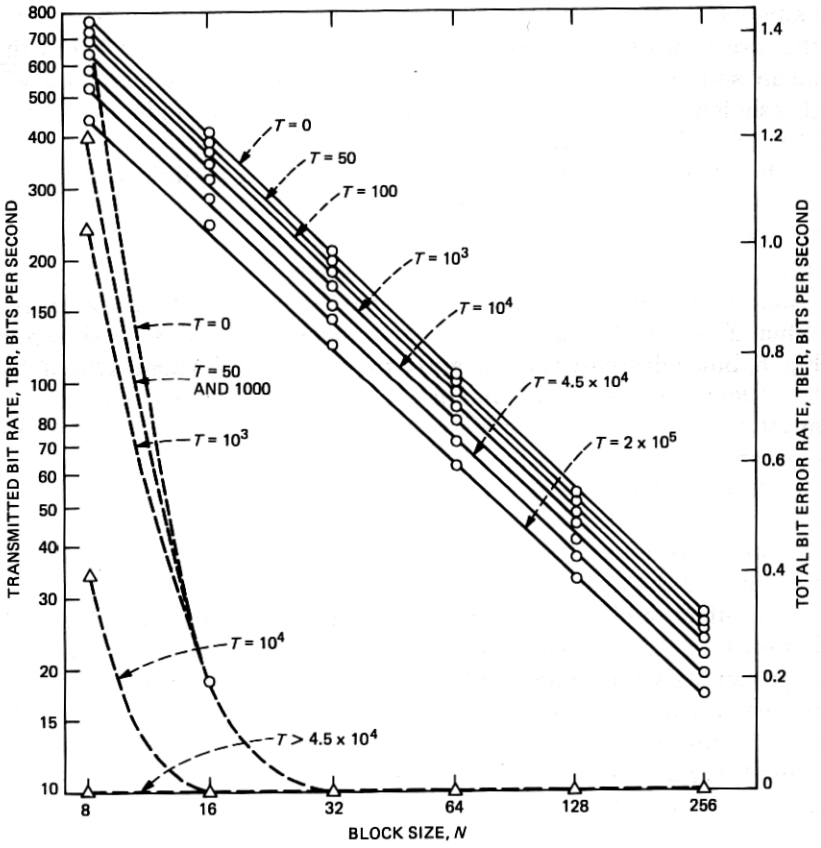

Fig. 8—Variation of data transmission efficiency (η) and data transmission deficiency (λ), as a function of block size (N) for different values of T, $K = 0.5$.

in the selection of blocks to convey data is based on whether the correlation coefficient ρ is above K. Increasing σ_{ni}^2 causes the speech to be decorrelated; therefore, less blocks have $\rho > K$, and consequently η decreases. Increasing T means that fewer blocks fulfill the condition that y of eq. (5) is a logical 1. The blocks discarded are generally those containing low-level speech, and it is these blocks that experience greatest decorrelation. Thus, as σ_{ni}^2 increases, η remains constant as the decorrelative effect is masked by the value of T. When σ_{ni}^2 approaches T, blocks not rejected because their mean value is $>T$ are now abandoned because of ρ being too small due to the decorrelation. Consequently, η versus σ_{ni}^2 is no longer a constant, and η coalesces with the $T = 0$ curve as σ_{ni}^2 is further increased. This occurs because the controlling factor in block rejection is now the correlation criterion. For $\sigma_{ni}^2 > T$, η decreases at approximately 6.8 percent per decade

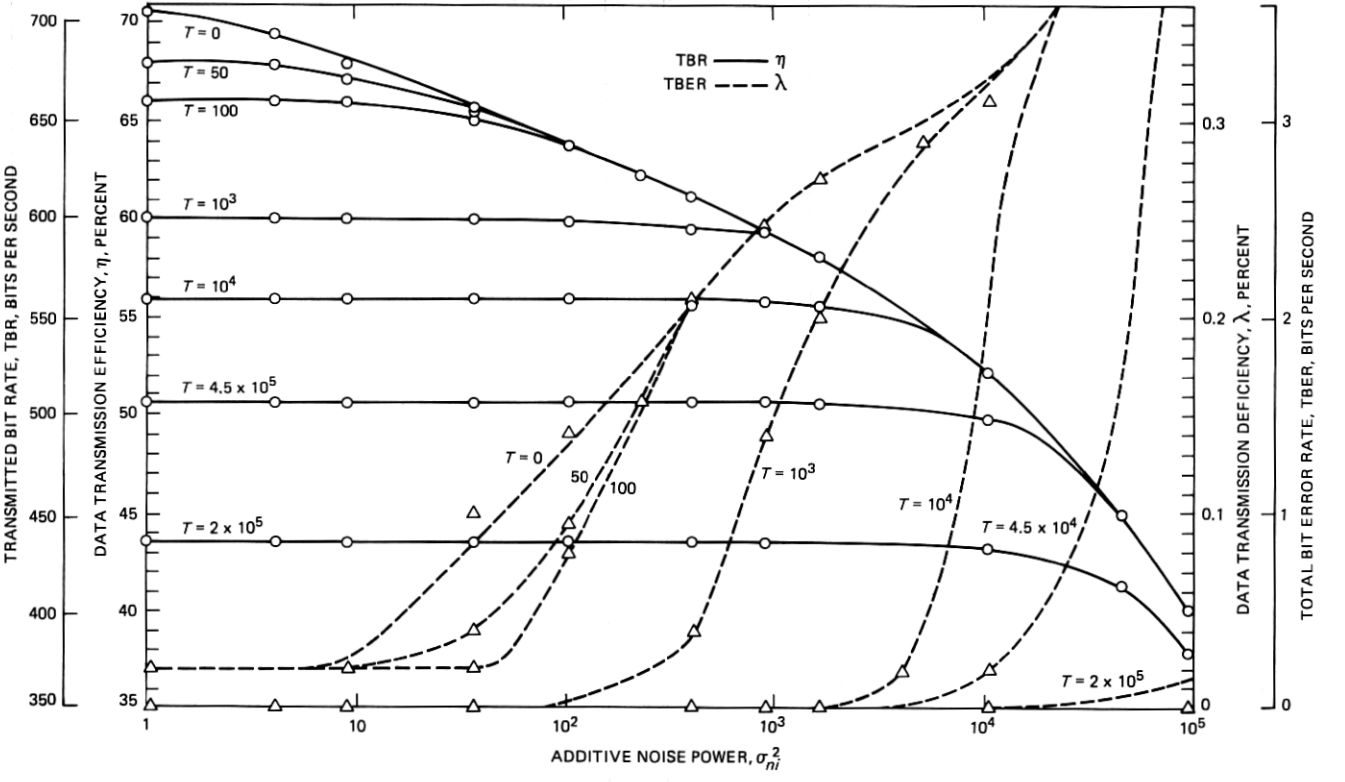

Fig. 9—Effect of additive background noise power (σ_{ni}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function of threshold T. $K = 0.6$, $N = 8$.

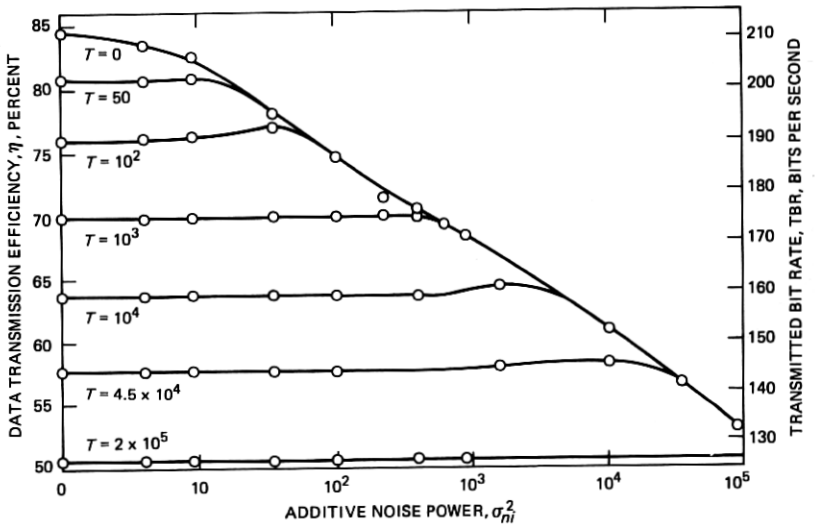

Fig. 10—Effect of additive background noise power (σ_{ni}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function of threshold T. $K = 0.5$, $N = 32$, $\lambda = 0$.

increase in σ_{ni}^2 when N is 32, and at a rate approaching this value when N is 8.

The data transmission deficiency λ and the TBER was found to be zero for the case of $N = 32$, over the range of power levels shown in Fig. 10. However, when the block size was 8, data errors occurred. The variation of λ and TBER with σ_{ni}^2 for different values of T is shown in Fig. 9. The figure demonstrates that data errors can be avoided for $\sigma_{ni}^2 < 10^4$ by setting T to 2×10^5, although the data transmission efficiency falls to 43.5 percent, i.e., TBR = 435 b/s.

4.3 Noisy channel

When no noise was added to the speech signal, but the channel was noisy with additive channel noise power σ_{nc}^2, η was uneffected. However, λ increased due to blocks that did not contain data ($\sigma_x^2 < T$) but had their power increased to $\sigma_x^2 + \sigma_{nc}^2 \geq T$, and if $\hat{\rho}$ or $\hat{\rho}_s$ exceeded K, data errors ensued. The variation of λ and TBER with σ_{nc}^2 for various values of T is displayed in Figs. 11 and 12, for $N = 8$ and 32, respectively. As λ and TBER had zero values for large values of T when $N = 32$, we present a zero line in Fig. 12, and the lines from this base to the other points on the curve are dotted. Notice in Fig. 12 that no data errors were recorded over the entire range of σ_{nc}^2 when $T = 2 \times 10^5$, and from Figure 10 this value of T corresponded to $\eta = 50.5$ percent. Thus, by using $N = 32$, and a background noise power and additive channel noise power up to 10^5, i.e., up to 10.4 dB below MS_x,

we found that 126 b/s can be transmitted without error. When $N = 8$, $T = 2 \times 10^5$, and both types of noise are present up to 10^3, TBR = 435 b/s, but TBER is no longer zero, having a value between 0.2 to 3.5 b/s. The various combinations of TBR and TBER can be deduced from Figs. 9 through 12.

V. SPEECH TRANSMISSION PERFORMANCE

Emphasis has been given to data transmission because we wanted to investigate if it could be reliably achieved using speech as a carrier signal. In the previous section, we presented results showing that it was possible to transmit data without transmission errors, and consequently the recovered speech signal was unimpaired by conveying the data. However, we have also observed that the data rate can be increased if bit errors can be tolerated. Thus, we now address the problem of how the bit errors affect the recovered speech signal, and specifically ask, How serious is the degradation of speech quality and intelligibility when the total TBER approaches the maximum values found in our experiments?

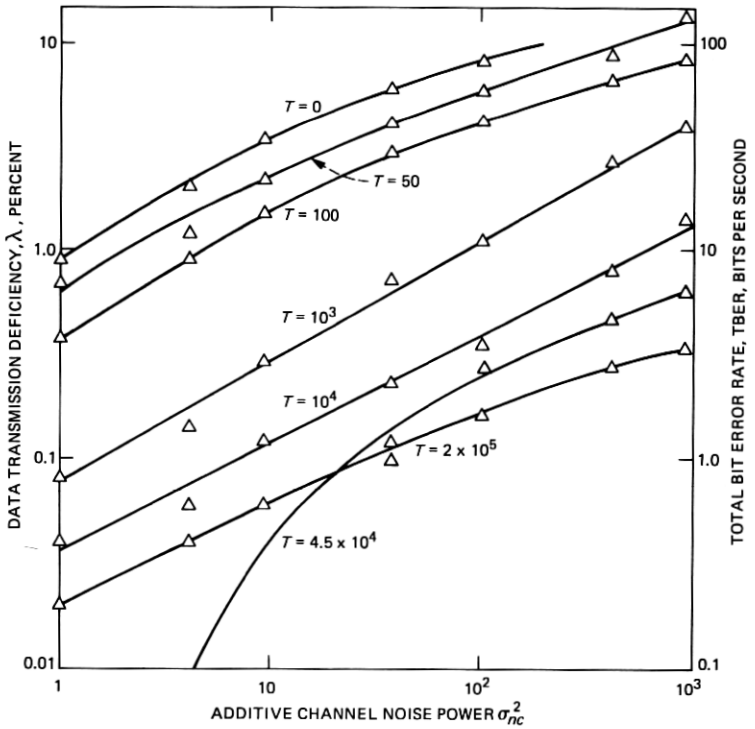

Fig. 11—Effect of additive channel noise power (σ_{nc}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function threshold T. $K = 0.6$, $N = 8$.

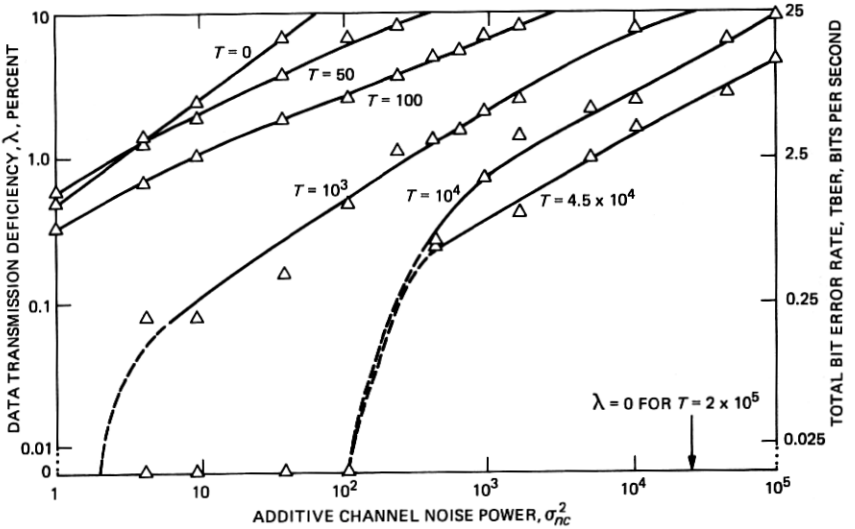

Fig. 12—Effect of additive channel noise power (σ_{nc}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function of threshold T. $K = 0.5$, $N = 32$, $\eta = 63.7$.

5.1 *Objective measure*

To provide an objective measure of the degradation of the recovered speech signal that accrues solely from the effect of data errors and not from the presence of additive background or channel noise, we elected to use segmental s/n (SEG-S/N). This ratio is used[5,6] as an objective measure because its value corresponds more closely to perceived quality than those of conventional s/n measurements, i.e., those using the ratio of mean square signal power to mean square noise power, determined over the duration of the entire signal. The reason for this resides in the computation of SEG-S/N, which is performed as follows. The input speech sequence $\{x_k\}$ is divided into contiguous blocks of 128 samples, i.e., into periods of 16 ms. Only those blocks, for example, M, whose rms value exceeds Γ dB (here -60 dB) relative to the peak value, have their s/n calculated. The s/n for the jth block is

$$\mathrm{s/n}_j = 10 \log_{10} \left\{ \frac{\sum_{i=1}^{128} x_{128j+i}^2}{\sum_{i=1}^{128} (x_{128j+i} - \hat{x}_{128j+i})^2} \right\}$$

$$j = 1, 2, \cdots, M, \tag{14}$$

where $\{\hat{x}_k\}$ is the recovered speech sequence. There is no point in considering $s/n_j < -10$ dB, or $> +80$ dB, because the speech quality is not perceived worse or better than -10 dB or $+80$ dB, respectively. Hence,

$$s/n_j = -10 \text{ dB for } s/n_j \leq -10 \text{ dB}$$

and

$$s/n_j = +80 \text{ dB for } s/n_j \geq +80 \text{ dB} \tag{15}$$

for $j = 1, 2, \cdots, M$. The number of data blocks of length N contained in the s/n_j calculation block size of 128, decreases from 16 to 0.5 as N is increased from 8 to 256, respectively. For a given TBER, the effect of increasing N is to cause the number of blocks not having the maximum s/n_j of 80 dB to decrease, but the decrease in s/n_j is more substantial in those blocks of 128 samples where erroneous scrambling occurred with all the samples, compared to those blocks where only 8 samples were erroneously scrambled.

The segmental s/n is computed as the average of s/n_j; $j = 1, 2, \cdots, M$, namely,

$$\text{SEG-S/N} = \frac{1}{M} \sum_{j=1}^{M} s/n_j . \tag{16}$$

The Γ threshold enables us to ignore blocks of low-level speech in the calculation of SEG-S/N. Even if these blocks are erroneously scrambled at the receiver output, their removal from the calculation is justified on the basis that the effect on such a low-level sound is imperceptible.

Let us now consider the case of speech conveying data through an ideal channel without the introduction of data errors. Here SEG-S/N is 80 dB, as s/n_j, for all j, is forced to 80 dB. When data errors occur, the SEG-S/N does not fall greatly below 80 dB, implying that the degradation in speech quality because of the presence of data errors is small. Perceptual observations substantiate this implication, confirming our decision to use SEG-S/N as an objective measurement of performance.

When the input speech is contaminated by statistically independent background noise, the recovered speech at the receiver is the sum of the original speech and noise signals, provided there are no data errors. Whenever a data error occurs, both the speech and the noise in the block are scrambled. In the case of additive channel noise, and for no data errors, the recovered speech is, again, the sum of the original speech and the noise signals, except when the data is a logical 1 when the output noise signal is scrambled. The effect of data errors, excluding states B and F, is to cause the output signal to be the sum of the scrambled original speech signal and the scrambled noise signal. Perceptually, the scrambled noise signal is the same as the original random

noise signal; therefore, its effect must be removed in evaluating the loss of SEG-S/N because of data errors. This is achieved by noting those blocks where data errors occurred in a noisy environment, and then scrambling the original speech in those same blocks to give the sequence $\{\hat{x}_k\}$ used in eq. (14). By this method, we are able to separate the distortion in the output signal, caused by the additive noise, from the effect of the data errors that were precipitated by this noise.

5.1.1 Objective results

The occurrence of a bit error results in a block of speech samples at the output of the receiver being erroneously scrambled. Clearly, if this happens to a block containing high amplitude samples, significant distortion ensues. From Section IV we have seen that the reduction or elimination of data errors can be achieved by increasing T. However, this is not our purpose here. We wish to generate data errors and observe their effect on the recovered speech. Consequently, in selecting conditions to illustrate the reduction in SEG-S/N caused by data errors, we have opted for $T = 0$. Table II shows the SEG-S/Ns for some of the worst data-error conditions shown in Figs. 5 to 12, plus a high data-error case when $N = 256$. These conditions were selected to show that in spite of the TBER values being unacceptably high for most data communications systems, the effect of the data errors on speech quality is small. Indeed, SEG-S/N remains above 66 dB for all the conditions depicted in Table II. Therefore, we refrain from presenting detailed measurements of speech distortion that is barely perceptible. However, we do discuss the error conditions, but cannot make general deductions from the few entries in the table, particularly as there is not a consistent theme. For example, for the noisy channel there are three values of N, but they each have a different K, so comparisons must be tempered with caution.

In Table I we have included the recovered speech status at the receiver for each of the data-error states. Two states, B and F, do not cause the recovered speech signal to have the samples in the erroneous blocks scrambled. When additive channel noise is present, error states D and F occur more frequently than the other states. Thus, the distortion in the output speech results mainly from state D, i.e., when a transmitted logical 1 is ignored. As states D and F are likely to occur with approximately the same probability (see Table II), the error rate in terms of distorting the recovered speech, can be considered to be reduced by a factor of two. However, state D is associated with a loss of a data signal caused by the channel noise increasing the correlation of a block of speech samples. Because data was transmitted, the speech can be voiced (although $T = 0$ for Table II) in which case the speech distortion may be substantial.

Table II—SEG-SN for some high data-error conditions, T = 0

Condition	Block Size N	Corr. Threshold K	Additive Noise Power σ_n^2	Occurrence of Data-Error States						Data Transmission Deficiency λ, %	TBER b/s	Data Transmission Efficiency η, %	TBR b/s	SEG-S/N dB
				A	B	C	D	E	F					
Ideal channel, no additive noise	8	0.2	ZERO	154	0	0	0	0	0	3.17	31.7	89.7	897	72.2
Noisy channel	8	0.6	918	6	32	1	302	1	286	12.8	128	80.0	800	66.7
Ideal channel, additive background noise	8	0.6	9.18×10^4	34	0	0	0	0	0	0.47	4.7	40.0	400	76.1
Ideal channel, no additive noise	32	0.2	ZERO	36	0	0	0	0	0	2.96	7.4	91.3	228	74.5
Noisy channel	32	0.5	918	0	0	0	99	0	85	15.5	38.8	84.5	211	75.6
Noisy channel	256	0.2	4.3×10^3	1	0	0	10	0	10	14.6	4.56	93.4	29	77.4

The data errors resulting from the ideal channel, with or without additive background noise, cause state A to apply, as shown by the examples in Table II. Although the output speech blocks are erroneously scrambled every time a data error occurs, the distortion is confined to blocks that often contain unvoiced sounds as the data error is the result of no data being transmitted, but a logical 1 being falsely generated.

Waveforms for the worst condition shown in Table II, namely, additive channel noise, $N = 8$, $T = 0$, $K = 0.6, \sigma_{nc}^2 = 918$, are displayed in Fig. 13. The high amplitude signal levels are seen to be substantially unaffected by the data errors whose effects are often immersed in the channel noise, and are therefore not perceptibly annoying.

5.2 Informal listening experiences

Informal listening tests were performed for the conditions listed in Table II. The recovered two sentences of speech, stripped of noise, with blocks of speech erroneously scrambled where data errors occurred, suffered only minor distortions. For the ideal channel, minor distortions resembling a "sshing" sound, occurred on three occasions for the cases of $N = 8$ and 32. A quiet noise, like additive white noise, was perceived for $N = 8$ when either background, or channel noise (the worst condition) were present. For the noisy channel condition, $N = 32$ produced the effect of barely perceptible scratches, while $N = 256$ yielded the least distortion, where the degradations were reminiscent of barely audible metallic clicks.

When the noisy output signal containing the effects of data errors was compared to the original speech plus additive noise, the effect of the data errors was imperceptible in the case of the substantial additive input noise power $\sigma_{ni}^2 = 9.18 \times 10^4$, N being 8. The effect of unwanted scrambling when the channel was noisy ranged from barely perceptible, $N = 256$ and 32, to nonannoying crackles when $N = 8$ and σ_{nc}^2 was only 918.

The conclusion is that for the data-error rates of practical significance, the degradation in speech quality is insignificant.

The transmitted signal sounded like distorted speech, plus white noise for the case of $N = 8$, and an ideal channel. The effect of additive background or channel noise was to reduce the fatiguing effects, as if the distortion had been removed from the speech and the background noise increased. When $N = 32$, the channel ideal, the distortion was increased as this block size corresponds to 4 ms, approximately half a pitch period. The speech sounded as if speaking and gargling were being performed simultaneously. The act of adding noise marginally reduced listening fatigue. The scrambled signal was found to be just intelligible when $N = 256$.

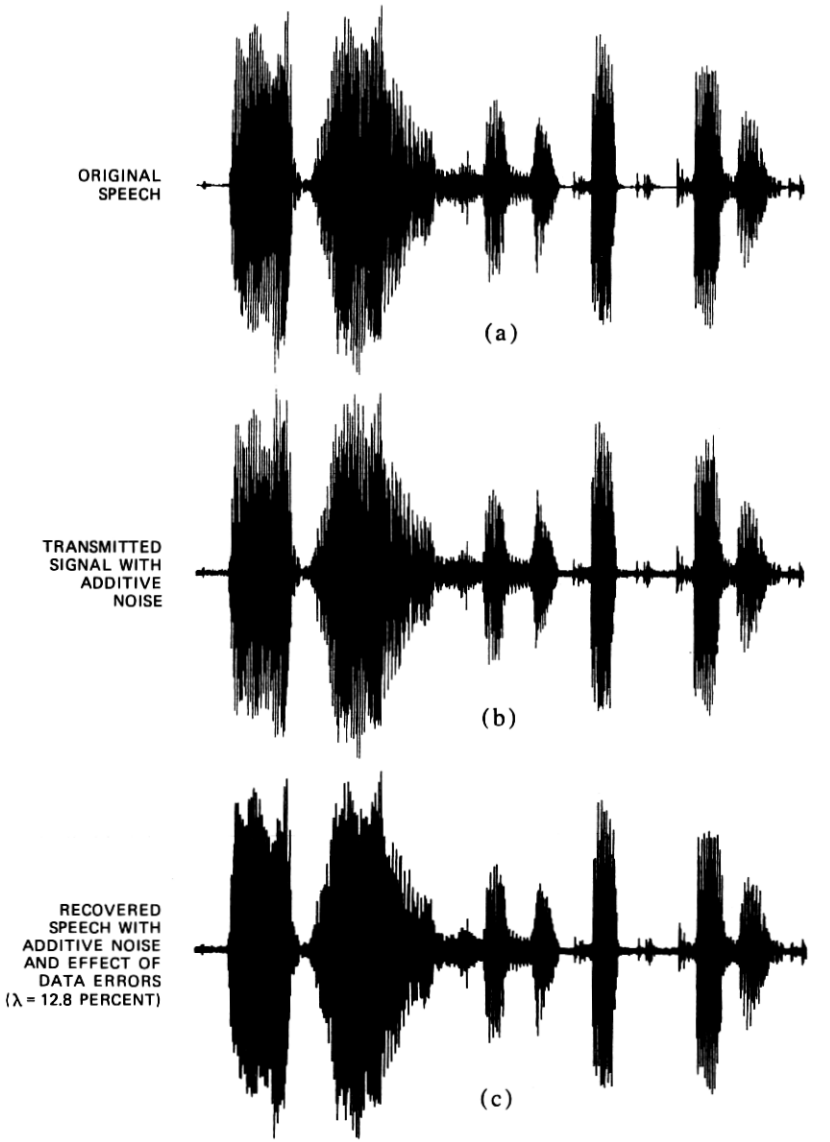

Fig. 13—Effect of additive channel noise. (a) Original speech. (b) Transmitted signal with additive channel noise having $\sigma_{nc}^2 = 920$. (c) Recovered speech having data errors ($\lambda = 12.8$ percent) and additive noise. $N = 8$, $T = 0$, $K = 0.6$.

VI. DISCUSSION

We started by enunciating a principle: that data could be transmitted by making it the scrambling key, and casting the receiver in the role of code breaker. Every time the receiver guesses the key, it obtains the

correct data and the correct speech. The speech is made an unwitting data carrier, while the data gets a free ride. The implications of this concept are considerable. Continuous users, or providers, of telephone traffic can, at the expense of additional terminal equipment, surreptitiously transmit teleprinter data, with the proviso that the bandwidth of the speech signal and the block size are appropriate for the channel bandwidth.

To demonstrate the principle, we wanted a scrambling technique that was easy to implement, and where the bandwith of the scrambled speech was not larger than that of the original speech. Frequency inversion scrambling aptly fulfilled these requisites, where the scrambling is achieved by merely altering the polarity of every other speech sample. We have shown that by using this form of scrambling, it is possible to transmit speech and data simultaneously, and to receive the data without errors and the speech without distortion, even in the presence of additive-channel and background noise. Provided some data errors can be tolerated, the data rate can be substantially increased as shown in Figs. 6–12. Even at high data-bit rates the distortion in the speech was found to be minimal, as the results in Table II indicate.

We have not presented results for dispersive channels, although we did do some experiments. The effect of such channels was to significantly alter the PDF of the correlation coefficient of the received signal compared to that of the transmitted signal. The power in the blocks of speech was also changed by the dispersive properties of the channel. As the data detection procedure is based on a measurement of power and correlation in a block of N samples, where the correlation is usually the most important factor, the dispersive channel results in an unacceptably high TBER. Thus, in the presence of dispersive channels, equalization of the channel must be performed.

The speech used in our experiments were two sentences whose waveforms are displayed in Fig. 3. Therefore, the results will differ when other sentences are used, but not significantly, as the sentences used consisted of over thirty-eight thousand samples. The system proposed here is for conveying data on speech or short silences. When prolonged silences occur, we envisage data being transmitted by conventional modern techniques.

The basic principle established, the way forward is to find scrambling methods that will be easier to break with certainty, and will operate via dispersive channels without the necessity of channel equalization.

VII. ACKNOWLEDGMENT

The authors thank D. J. Goodman for his constructive criticism of this work.

REFERENCES

1. D. Kahn, *The Code-Breakers*, New York: Macmillan, 1967.
2. G. Guanella, "Automatic Speech Scrambling," Brown, Boveri & Company Pub. CH-E7.30038.2E.
3. S. C. Kak and N. S. Jayant, "On Speech Encryption Using Waveform Scrambling," B.S.T.J., *56,* No. 5 (May–June 1977), pp. 781–808.
4. N. S. Jayant et al., "A Comparison of Four methods for Analog Speech Privacy," IEEE Trans. Commun., *COM-29,* No. 1 (January 1981), pp. 18–23.
5. B. McDermott, C. Scagliola, and D. J. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM," B.S.T.J., *57,* No. 5 (May–June 1978), pp. 1597–618.
6. J. M. Tribolet et al., "A Comparison of the Performance of Four Low-Bit-Rate Speech Waveform Coders," B.S.T.J., *58,* No. 3 (March 1979), pp. 699–712.