Log Ratio of Entropy Powers

Jerry D. Gibson
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, California
gibson@ece.ucsb.edu

***Abstract*—In his landmark 1948 paper [1], Shannon defined what he called the derived quantity of entropy power, also called entropy rate power, to be the power in a Gaussian white noise limited to the same band as the original ensemble and having the same entropy. He then used the entropy power in bounding the capacity of certain channels and for specifying a lower bound on the rate distortion function of a source. Kolmogorov [2] and Pinsker [3] derived an expression for the entropy rate power of a discrete-time stationary random process in terms of its power spectral density. The entropy power inequality also plays an important role in multiterminal information theory. These have been the major applications of entropy power in the last (almost) 70 years. We reconsider entropy rate power and use the log ratio of entropy powers to analyze the performance of tandem communications and signal processing systems in terms of the change in mutual information with each stage. We also examine ways to calculate or approximate the entropy rate power when performing the analyses.**

***Index Terms*—entropy rate power, cascade signal processing, mutual information, tandem communications systems**

I. Introduction

Having been introduced by Shannon in his original, landmark 1948 paper, *The Mathematical Theory of Communication* [1], entropy rate power is as fundamental a concept in Shannon Information Theory as mutual information (then called average mutual information). Entropy rate power shows up in some basic bounds on communications and compression performance, also first suggested by Shannon in 1948, and forms the basis for some interesting inequalities [4]. However, unlike many other quantities in Shannon theory, entropy rate power is not a well-studied, widely employed concept. We examine entropy rate power from several different perspectives that reveal new possibilities, and we define a new quantity, the log ratio of entropy powers, that serves as a new tool in the analysis of tandem communications systems and in the information theoretic study of cascade signal processing operations. Gibson and Mahadevan [5] have used the log ratio of entropy powers to derive and extend the interpretation of the log likelihood spectral distance measure from signal processing.

The concept of entropy power and entropy rate power are defined in Sec. II and the well known results for entropy power in terms of the power spectral density of a Gaussian process are summarized. Methods for calculating the entropy power are outlined in Sec. III, including how the entropy power might be approximated. The equivalence of the entropy power and the minimum mean squared one step ahead prediction error for Gaussian autoregressive sequences is discussed in Sec. IV. Section V sets up a basic cascade signal processing problem and the well known inequalities for mutual information in the signal processing chain are stated. Inequalities in terms of the entropy power that follow from the results in Sec. V are given. The fundamental new quantity, the log ratio of entropy powers, is developed in Sec. VI, explicitly stating its relationship to the changes in mutual information as a signal progresses through the signal processing chain. Section VII shows when our new results align with those obtained earlier by Messerschitt [6]. Section VIII provides an analysis of a problem considered by Wolf and Ziv [7] in terms of the log ratio of entropy powers. Conclusions are provided in Sec. IX.

II. Entropy Power/Entropy Rate Power

In his landmark 1948 paper [1], Shannon defined what he called the *derived quantity* of entropy power (also called entropy rate power) to be the power in a Gaussian white noise limited to the same band as the original ensemble and having the same entropy. He then used the entropy power in bounding the capacity of certain channels and for specifying a lower bound on the rate distortion function of a source. Within the context of calculating and bounding the rate distortion function of a a discrete-time stationary random process, Kolmogorov [2] and Pinsker [3] derived an expression for the entropy rate power in terms of its power spectral density.

Given a random variable X with probability density function $p(x)$, we can write the differential entropy

$$h(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx \tag{1}$$

where X has the variance $var(X) = \sigma^2$. Since the Gaussian distribution has the maximum differential entropy of any distribution with mean zero and variance σ^2 [4],

$$h(X) \leq \frac{1}{2} \log 2\pi e \sigma^2 \tag{2}$$

from which we obtain

$$Q = \frac{1}{(2\pi e)} \exp 2h(X) \leq \sigma^2 \tag{3}$$

where Q was defined by Shannon to be the *entropy power* associated with the differential entropy of the original random variable [1]. In addition to defining entropy power, this equation shows that the entropy power is the *minimum variance* that can be associated with the not-necessarily-Gaussian differential entropy $h(X)$.

Note that Eq. (3) allows us to calculate the entropy power associated with a given entropy. For example, if the random variable X is Laplacian [4] with parameter λ, then $h(X) = \ln(2\lambda e)$ and we can substitute this into Eq. (3) and solve for the entropy power Q. See Sec. III.

For an n-vector $\mathbf{X}$ with probability density $p(x^n)$, and covariance matrix $K_{\mathbf{X}} = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T]$, we have that

$$h(\mathbf{X}) \leq \frac{1}{2} log[(2\pi e)^n |K_{\mathbf{X}}|] \tag{4}$$

from which we can construct the vector version of the entropy power as

$$Q_{\mathbf{X}} = \frac{1}{(2\pi e)^n} \exp 2h(\mathbf{X}) \leq |K_{\mathbf{X}}|. \tag{5}$$

We will have the occasion to study pairs of random vectors $\mathbf{X}$ and $\mathbf{Y}$ where we use the vector $\mathbf{Y}$ to form the best estimate of $\mathbf{X}$. If $K_{\mathbf{X}|\mathbf{Y}}$ is the covariance matrix of the minimum mean squared error estimate of $\mathbf{X}$ given $\mathbf{Y}$, then we have

$$h(\mathbf{X}|\mathbf{Y}) \leq \frac{1}{2} log[(2\pi e)^n |K_{\mathbf{X}|\mathbf{Y}}|] \tag{6}$$

and from which we can get an expression for the conditional entropy power, $Q_{\mathbf{X}|\mathbf{Y}}$,

$$Q_{\mathbf{X}|\mathbf{Y}} = \frac{1}{(2\pi e)^n} \exp h(\mathbf{X}|\mathbf{Y}) \leq |K_{\mathbf{X}|\mathbf{Y}}|. \tag{7}$$

So, $Q_{\mathbf{X}|\mathbf{Y}}$ is upper bounded by the determinant of the conditional error covariance matrix, $|K_{\mathbf{X}|\mathbf{Y}}|$. We have equality in Eqs.(2) - (7) if the corresponding random variables or random vectors are Gaussian.

Often our interest is in investigating the properties of stationary random processes. Thus, if we let $\mathbf{X}$ be a stationary continuous-valued random process with samples X^n =$[X_i, i = 1, 2, ..., n]$, then the differential entropy rate of the process $\mathbf{X}$ is [8]

$$\overline{h} = \lim_{n\to\infty} \frac{1}{n} h(X^n) = \lim_{n\to\infty} h(X_n|X^{n-1}) \tag{8}$$

We assume that this limit exists in our developments and we drop the overbar notation and use $h = \overline{h}$. Using the entropy rate in the definition of entropy power yields the nomenclature entropy rate power.

If we now consider a discrete-time stationary Gaussian process with correlation function $\phi(k) = E[X_j X_{j+k}]$, the periodic discrete-time power spectral density is defined by

$$\Phi(\omega) = \sum_{-\infty}^{\infty} \phi(k) \exp(j\omega k) \tag{9}$$

for $|\omega| \leq \pi$. We know that an n-dimensional Gaussian density with correlation matrix Φ_n has the differential entropy $h(\mathbf{X})$ = $(n/2)\log(2\pi e|\Phi_n|^{1/n})$. Then, the entropy rate power Q can be found from [9], [10]

$$\log Q = \lim_{n\to\infty} \log |\Phi_n|^{1/n} \tag{10}$$

which yields [9], [10]

$$Q = \exp[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \Phi(\omega) d\omega] \tag{11}$$

as the entropy rate power.

In rate distortion theory, Q can be used to write a lower bound on the rate distortion function. In particular, Shannon showed that the rate distortion function for a source Y is bounded as

$$\frac{1}{2}\log \frac{Q_Y}{D} \leq R_Y(D) \leq \frac{1}{2}\log \frac{\sigma_Y^2}{D}, \tag{12}$$

where σ_Y^2 is the variance of the source and Q_Y is the entropy power of the source. The lower bound involving the entropy rate power is one of the most quoted results in rate distortion theory.

The fact that Q can be used to obtain a lower bound on the rate distortion function for the mean squared error distortion measure relates to the fact that Q is the minimum variance that can be associated with the source Y differential entropy.

III. Calculating the Entropy Power

Before we start considering several examples, we need to think about how to calculate the required entropy powers. For a given memoryless source Z with differential entropy $h(Z)$ and variance $Var(Z)$, we can find the corresponding entropy power by setting

$$h(Z) = \frac{1}{2}\log(2\pi e Q) \tag{13}$$

and solving for Q. For example, if Z is a Laplacian random variable, then $h(Z) = ln(2\lambda e)$ where $Var(Z) = 2\lambda^2$, so

$$Q = \frac{1}{2\pi} \exp(1 + 2\lambda) \tag{14}$$

If the distribution is not known, a goodness-of-fit test can be performed to find the closest match and then the entropy corresponding to that distribution can be used as in Eq. (13) to find the entropy power.

If the signals are Gaussian, Eq. (6) is satisfied with equality and then the conditional error variance $|K_{\mathbf{X}|\mathbf{Y}}|$ is the entropy power. Irrespective of the distribution, when the minimum mean squared error estimator is available, since the entropy power is the minimum variance that can be associated with a given distribution (from the estimation counterpart to Fano's Inequality [4]), then the corresponding minimum mean squared error can serve as an approximation to the entropy power since it is an upper bound. For convenience, we may try to simplify the calculations even further and use the (not minimum) mean squared error at each stage as the entropy power, but the resulting calculations can be grossly in error. How inexact will depend on the signals themselves.

IV. Minimum Mean Squared Prediction Error in Time Series Analysis

Perhaps surprisingly, Eq. (11) plays a major role in time series analysis. It is the minimum mean squared one-step prediction error for autoregressive (AR) processes, even if they are not Gaussian [9], [11], [12]! Other than appearing in the lower bound of $R(D)$ in Eq. (12), this fact has not produced an impact on information theory or rate distortion theory.

One place the fact that Eq. (11) is the minimum mean squared one-step prediction error for AR processes, even non-Gaussian processes, is useful is on upper bounding the entropy power for autoregressive processes. This is particularly interesting since not only do we have Eq. (11) as a way to calculate the entropy rate power for Gaussian processes, but there is a vast literature on finding the AR coefficients that minimize the one step ahead mean squared prediction error [13]–[16] in the field of linear prediction for speech and seismic signal processing.

Two things thus follow. If the AR or linear prediction coefficients are known, it is straightforward to find the minimum mean squared prediction error [16], [17]. Further, if we fit an AR model to a time series, we can find the minimum mean squared prediction error and hence upper bound the entropy rate power that corresponds to that model. If the time series is in fact AR, we can iteratively increase the model order, calculate the AR or linear prediction coefficients for the iterated model order, and then find the best fit, from which we can calculate the upper bound to the entropy rate power. If the time series is not AR, then we can find the best approximation and its corresponding minimum mean squared error, which is again an upper bound on the entropy power for that best fit model.

V. Cascaded Signal Processing

Figure 1 shows a cascade of N signal processing operations with the Estimator blocks at the output of each stage as studied by Messerschmitt [6]. He used the conditional mean at each stage and the corresponding conditional mean squared errors to obtain a representation of the distortion contributed by each stage. We analyze the cascade connection in terms of information theoretic quantities, such as mutual information, differential entropy, and entropy rate power. Similar to Messerschmitt, we consider systems that have no hidden connections between stages other than those explicitly shown. Therefore, we conclude directly from the Data Processing Inequality [4] that

$$I(X;Y_1) \geq ... \geq I(X;Y_{N-1}) \geq I(X;Y_N) \geq I(X;\widehat{X}) \quad (15)$$

Since $I(X;Y_n) = h(X) - h(X|Y_n)$, it follows from Eq. (15) that

$$h(X|Y_1) \leq ... \leq h(X|Y_{N-1}) \leq h(X|Y_N) \leq h(X) \quad (16)$$

We do not consider the special cases when the differential entropy is negative, which in our applications can be avoided by normalization.

For the optimal estimators at each stage, the basic Data Processing Inequality also yields

$$I(X;Y_n) \geq I(X;\widehat{X}_n) \quad (17)$$

and thus

$$h(X|Y_n) \leq h(X|\widehat{X}_n). \quad (18)$$

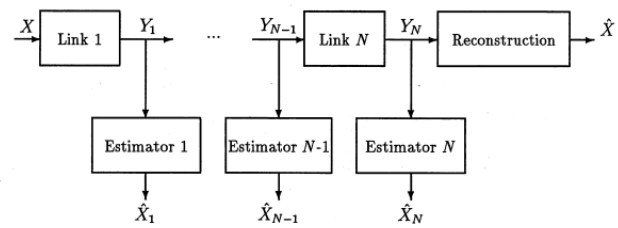

Fig. 1. N-Link System Block Diagram (Adapted from [6])

These are the fundamental results that additional processing cannot increase the mutual information.

Now we notice that the series of inequalities in Eq. (16) along with the entropy power expression in Eq. (3) gives us the series of inequalities in terms of entropy power at each stage in the cascaded signal processing operations

$$Q_{X|Y_1} \leq ... \leq Q_{X|Y_{N-1}} \leq Q_{X|Y_N} \leq Q_X \quad (19)$$

Equation (18) also allows us to write that

$$Q_{X|Y_n} \leq Q_{X|\widehat{X}_n} \quad (20)$$

In the context of Eq. (19), the notation $Q_{X|Y_n}$ denotes the minimum variance when reconstructing an approximation to X given the sequence at the output of stage n in the chain. Equivalently, we can think of the entropy power $Q_{X|Y_n}$ as the variance of the error for the optimum reconstruction of X given Y_n when the Shannon lower bound for the squared error difference distortion measure is satisfied with equality [9]. In the following, we build on these ideas to obtain new results in signal processing.

It is important to notice that although the definition of entropy power uses the expression for the differential entropy of a Gaussian random variable, to this point, there have been no assumptions on the differential entropy, and hence, on the distributions of the random variables, at each stage in the signal processing chain. The entropy power is unique in that it offers a connection to differential entropy, mutual information, and minimum mean squared error, which we exploit in the following.

VI. Log Ratio of Entropy Powers

We can use the definition of the entropy power in Eq. (3) to express the logarithm of the ratio of two entropy powers in terms of their respective differential entropies as

$$\log \frac{Q_X}{Q_Y} = 2[h(X) - h(Y)] \quad (21)$$

We can write a conditional version of Eq. (3) as

$$Q_{X|Y_n} = \frac{1}{(2\pi e)} \exp 2h(X|Y_n) \leq Var(X|Y_n) \quad (22)$$

and from which we can express Eq. (21) in terms of the entropy powers at successive stages in the signal processing chain as

$$\frac{1}{2} \log \frac{Q_{X|Y_n}}{Q_{X|Y_{n-1}}} = h(X|Y_n) - h(X|Y_{n-1}) \quad (23)$$

If we add and subtract $h(X)$ to the right hand side of Eq. (23), we then obtain an expression in terms of the difference in mutual information between the two stages as

$$\frac{1}{2}\log\frac{Q_{X|Y_n}}{Q_{X|Y_{n-1}}} = I(X;Y_{n-1}) - I(X;Y_n) \tag{24}$$

From the series of inequalities on the entropy power in Eq. (19), we know that both expressions in Eqs. (23) and (24) are greater than or equal to zero. It is important to note that Eqs. (23) and (24) are not based on a Gaussian assumption for the underlying processes.

We can relate the log ratio of entropy powers to the change in mutual information over the entire signal processing chain as follows. The entropy power of the input is Q_X and the conditional entropy power at the output $\widehat{X}$ is $Q_{X|\widehat{X}}$ so

$$\begin{aligned}\frac{1}{2}\log\frac{Q_{X|\widehat{X}}}{Q_X} =& \frac{1}{2}\log\frac{Q_{X|Y_1}}{Q_X} + \frac{1}{2}\log\frac{Q_{X|Y_2}}{Q_{X|Y_1}} + \ldots \\ &+\frac{1}{2}\log\frac{Q_{X|Y_n}}{Q_{X|Y_{n-1}}} + \frac{1}{2}\log\frac{Q_{X|\widehat{X}}}{Q_{X|Y_n}} \\ =& I(X;X) - I(X;Y_1) + I(X;Y_1) - I(X;Y_2) \\ &+ \ldots + I(X;Y_{n-1}) - I(X;Y_n) \\ &+ I(X;Y_n) - I(X;\widehat{X}) \\ =& I(X;X) - I(X;\widehat{X})\end{aligned} \tag{25}$$

We see that the loss in mutual information from input to output is as expected. If $\widehat{X} \cong X$, then the information loss will be low, but if $\widehat{X}$ is almost independent of the input X, then the loss in mutual information will be nearly complete. Inspecting this sequence of terms in Eq. (25), we see how to find the difference in mutual information between the input and any intermediate stage or between any pair of stages in the cascaded signal processing operations.

These results are new and extend the Data Processing Inequality by providing a new characterization of the information loss between stages in terms of the entropy powers of the two stages.

VII. Relationship to Messerschmitt's Results

With the Gaussian assumption, a specific connection to Messerschmitt's results can be obtained. Note that if all signals in the chain are Gaussian, then the conditional probability density at the output of each link, $p_{X|Y_n}(x|y_n)$, is Gaussian with mean $E(X|Y_n)$ and variance $Var(X|Y_n)$. Additionally, the conditional entropy at each stage is $h(X|Y_n) = \frac{1}{2}\log 2\pi e Var(X|Y_n)$, and $Q_{X|Y_n} = Var(X|Y_n)$. If we substitute these results into Eq. (16) or into Eq. (19), we obtain the relations

$$Var(X|Y_1) \leq \ldots \leq Var(X|Y_{n-1}) \leq Var(X|Y_n) \tag{26}$$

Notice that since all of the quantities are assumed to be Gaussian, Eq. (3) holds with equality and Eq. (26) follows directly from Eq. (19). Further, since Eq. (3) holds with equality for Gaussian processes, we have for all of the quantities at each stage that

$$\begin{aligned}&Q_X = Var(X) \\ &Q_{X|Y_n} = Var(X|Y_n), n = 1, ..., N \\ &Q_{\widehat{X}_n} = Var(\widehat{X}_n), n = 1, ..., N \\ &Q_{X|\widehat{X}_n} = Var(X|\widehat{X}_n), n = 1, ..., N;\end{aligned} \tag{27}$$

In his paper [6], Messerschmitt finds the minimum mean squared estimator at each stage, denoted $\widehat{X}_n$, and shows that

$$Var(\widehat{X}_1) \geq Var(\widehat{X}_2) \ldots \geq Var(\widehat{X}_N), \tag{28}$$

so that the variance of the *estimate* is decreasing in moving through the signal processing chain.

This is consistent with our results since at any stage n, we have that the relationship of the input variance, $Var(X)$ and the variance of the error $Var(X|Y_n)$ is given by

$$Var(X) = Var(\widehat{X}_n) + Var(X|\widehat{X}_n), \tag{29}$$

so that our Eq. (26) implies Eq. (28). The result in Eq. (29) can also be recognized as the relationship when the Shannon backward channel is achieved with equality [9].

Therefore, under the Gaussian assumption, our entropy rate power approach produces the minimum mean squared error at each stage just as in the conditional mean approach of Messerschmitt [6]. Messerschmitt notes that computing the conditional mean is difficult except in the case of Gaussian signals.

However, he did not consider forming the logarithm of the ratio of variances, and its' key relationships to the difference in differential entropies and mutual informations, and did not make an explicit connection to differential entropy and to mutual information as we are able to do so here with entropy rate power.

VIII. Wolf and Ziv Result

We also use our approach to analyze the system investigated by Wolf and Ziv [7] shown in Fig. 2. Wolf and Ziv extended the work of Dobrushin and Tsybakov [18], who had considered the same structure but under the assumption that the signals were Gaussian and that the source mapping is additive Gaussian noise, among others.

Wolf and Ziv consider the mean squared error distortion measure and for this system, the input source x is processed through a noisy Source Mapping and the optimum Encoder is to calculate the conditional mean of the input x given the output of the noisy mapping s and encode this conditional mean. The output of the channel is then input to the optimal decoder, which consists of the conditional mean estimate of μ given the channel output. This estimate is then delivered to the Final Destination through a noisy Receiver Mapping. For this optimal system, Wolf and Ziv find the total mean squared error averaged over an interval T, denoted as D_T, as

$$D_T = \overline{|\widehat{x} - x|^2} = \overline{|\mu - x|^2} + \overline{|v - \mu|^2} + \overline{|\widehat{x} - v|^2} \tag{30}$$

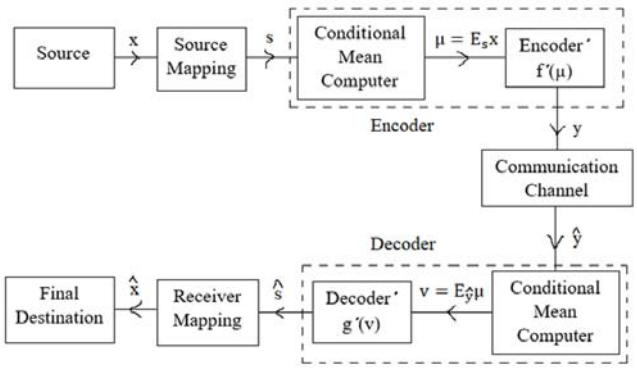

Fig. 2. Transmission of a Noisy Source to a Noisy Receiver (Adapted from [7])

To align with our notation, we let $X = x$, $Y_1 = s$, $Y_2 = \widehat{y}$, and $\widehat{X} = \widehat{x}$. We also note that $\mu = E(X|Y_1)$ and $v = E(\mu|\widehat{y}) = E[E(X|Y_1)|Y_2]$.

As a result, we can write in terms of our entropy power notation,

$$\begin{aligned} Q_{X|\widehat{X}} &= \overline{|\widehat{x} - x|^2} \\ Q_{X|Y_2} &= \overline{|\mu - x|^2} + \overline{|v - \mu|^2} \\ Q_{X|Y_1} &= \overline{|\mu - x|^2}. \end{aligned} \tag{31}$$

where we see that our conditional entropy power expressions capture the total error at each point in the signal processing chain. The mean squared error contributions for each step in the chain are differences in entropy powers.

In terms of the log ratios of entropy powers or the loss in mutual information at each stage for the Wolf and Ziv [7] results, we can now write

$$\begin{aligned} \frac{1}{2}\log\frac{Q_{X|\widehat{X}}}{Q_{X|Y_2}} =& \frac{1}{2}\log\frac{\overline{|\widehat{x} - x|^2}}{\overline{|\mu - x|^2} + \overline{|v - \mu|^2}} \\ =& I(X;Y_2) - I(X;\widehat{X}) \end{aligned} \tag{32}$$

for the last stage, and

$$\begin{aligned} \frac{1}{2}\log\frac{Q_{X|Y_2}}{Q_{X|Y_1}} =& \frac{1}{2}\log\frac{\overline{|\mu - x|^2} + \overline{|v - \mu|^2}}{\overline{|\mu - x|^2}} \\ =& I(X;Y_1) - I(X;Y_2). \end{aligned} \tag{33}$$

for the second stage, and

$$\begin{aligned} \frac{1}{2}\log\frac{Q_{X|Y_1}}{Q_X} =& \frac{1}{2}\log\frac{\overline{|\mu - x|^2}}{Q_X} \\ =& I(X;X) - I(X;Y_1). \end{aligned} \tag{34}$$

for the first stage.

As in Messerschmitt's example, the difference in mutual information for the first stage is actually just $h(X|Y_1) = h(x|s)$, so the first stage difference in mutual informations should not be considered a loss in mutual information.

IX. Conclusions

We present a new quantity, the log ratio of entropy powers, for investigating the changes in mutual information and differential entropy as signals progress through the cascade connection of signal processing operations. This work extends prior analyses of cascade connections that only considered the mean squared error contribution at each stage. Approaches to calculating the entropy power under different conditions are outlined. Explicit connections to the prior work of Wolf and Ziv, Dobrushin and Tsybakov, and Messerschimtt are developed. Note that by employing the entropy power, we have a set of assumptions intermediate to those of Dobrushin and Tsybakov [18] and Wolf and Ziv [7], and an information theoretic view of Messerschitt's analyses [6]. Further explorations of the log ratio of entropy powers, including fundamental theoretical relationships, the calculation of entropy power, and different examples, are underway.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 1948.

[2] A. N. Kolmogorov, "Theory of the transmission of information," *Proceedings of the Session on Scientific Problems of Automation*, 1956.

[3] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[5] J. D. Gibson and P. Mahadevan, "Log likelihood spectral distance, entropy rate power, and mutual information with applications to speech coding," *Entropy*, vol. 19, no. 9, p. 496, 2017.

[6] D. G. Messerschmitt, "Accumulation of distortion in tandem communication links," *IEEE Trans. on Information Theory*, vol. IT-25, no. 6, pp. 692–698, Nov. 1979.

[7] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 406–411, Jul. 1970.

[8] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

[9] T. Berger, *Rate Distortion Theory*. Prentice-Hall, 1971.

[10] T. Berger and J. D. Gibson, "Lossy Source Coding," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2693 –2723, Oct. 1998.

[11] U. Grenander and M. Rosenblatt, *Statistical analysis of stationary time series*. John Wiley & Sons, Almqvist & Wiksell, 1957.

[12] L. H. Koopmans, *The spectral analysis of time series*. Academic press, 1995.

[13] J. D. Markel and A. J. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013, vol. 12.

[14] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: Principles and Standards*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1998.

[15] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[16] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[17] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustic Society of America*, vol. 50, pp. 637–655, 1971.

[18] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, Sep. 1962.