SIGNAL PROCESSING III: Theories and Applications
I.T. Young et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

Adaptive Maximum Entropy Coding

N. Merhav and D. Malah
Electrical Engineering Department
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

Abstract

In this paper we analyze and examine a recently proposed waveform coding scheme based on maximizing the entropy of the transmitted bit-stream. The theoretical motivation for using this scheme is the fact that maximum entropy is a necessary condition for optimality of any coding scheme. A practical motivation is its simplicity and amenability to fast implementation. For stationary signals, a detailed analysis of the coder/decoder characteristics is presented. For non-stationary signals we propose an adaptation scheme which tracks slow temporal variations of some statistical parameters. A gain adaptation mechanism cancels the idle channel noise which cannot be removed by an ordinary A.G.C. The new adaptive system is found to overperform ADPCM, particularly for not too - highly correlated ($\rho < 0.8$) non-stationary Gaussian processes.

1. Introduction

In this paper, a new adaptive predictive waveform coding scheme is developed and examined. The scheme is based upon maximizing the first order entropy of the transmitted bit stream. This concept is proposed by E. Angel and L. Daigle in [1], who presented some results of a non-adaptive version of the system (for 1 and 2 bits/sample), for coding speech signals [2] and images [1]. These authors assume the input signal to be a stationary Gaussian process with a known covariance function. For this class of signals we found their proposed scheme to over perform DPCM significantly in a wide range of the correlation coefficient value. However, for non-stationary signals, the fixed scheme is not suitable. Moreover, low energy regions of large dynamic range signals, are reconstructed with very high level of idle channel noise. The non-linear nature and the implicit dependence of the encoder characteristics upon the input signal gain and correlation coefficient value, causes considerable difficulties in the adaptation task. Nevertheless, in this paper we propose approximations of these characteristics by explicit functions of the gain and the correlation coefficient, which enable adaptation to slow variations of these parameters. The suggested gain adaptation algorithm cancels the idle channel noise, which can not be removed by an ordinary A.G.C. However, the computational complexity required for adapting the correlation coefficient is greater than in classical ADPCM. The proposed adaptive predictive maximum entropy system is particularly suitable for coding non-stationary Gaussian processes with slowly varying covariance functions. For speech signals, the resulting quality and intelligibility are equivalent to CVSD for the 1 bit/sample version and to ADPCM for the 2 bit/sample system. In addition to simulation results, the presentation of an adaptation scheme, and the method for cancelling the limit cycles at low input signal amplitudes, an important contribution of this paper is a detailed analysis and presentation of the coding and decoding characteristics, not given previously in [1,2].

2. The Maximum Entropy Concept

The maximum entropy (ME) criterion has been proposed by Angel and Daigle [1,2]. As we see it, the motivation for selecting this criterion, is the fact that maximum entropy is a necessary attribute of an optimum data compression system. This follows from a simple consideration: Suppose we had such an optimum system having entropy less than its rate (in bits/symbol), then one could further compress the data with no additional distortion (by entropy coding), and consequently reduce the rate. It follows that the original rate was not minimum for the given allowed distortion. From the convexity property of the rate-distortion function, it follows that the distortion was not minimum for that rate, in contrast to the above assumption.

It can be shown [5] that for conventional waveform coding schemes such as DPCM, there are considerable difficulties in optimizing the parameters (in the MMSE sense) for low rates, because the true equations for solving the optimum predictor coefficients are highly non-linear. On the other hand, since it satisfies a necessary condition for optimality, the ME approach has the potential of obtaining an improved solution compared to a solution based on linearizing the non-linear equations.

3. System Description

In this section we review the scheme proposed by Angle and Daigle [1,2]. In order to obtain a convenient analytic solution, these authors [1,2] concentrate on the maximization of the conditional first order entropy and assume that the source is a Gaussian process with a known covariance function. Fig.1 depicts the transmitter and receiver for the rate of 1 bit/sample. The transmitter (a) is quite similar to the first order DPCM transmitter. However, in contrast to the DPCM transmitter, which is designed to minimize the energy of the residual, the suggested predictor (also 1st order), is designed in such a way that the output symbols are equally likely, given the information currently available to the predictor, i.e., e_n and y_n. In other

words, the following condition is to be satisfied:

$$\Pr\{e_{n+1}=1|e_n,y_n\}=\Pr\{e_{n+1}=0|e_n,y_n\}=\frac{1}{2} \quad (1)$$

In this way the maximum conditional first order entropy is ensured and consequently, statistical independence between successive bits. The receiver (b) consists of a similar predictor (in order to reconstruct y_n), and in addition an estimator of the input ($\hat{x}_n$) which utilizes the information currently available at the receiver; namely, e_n, y_n, e_{n-1} and y_{n-1}:

$$\hat{x}_n = E\{x_n|e_n,e_{n-1},y_n,y_{n-1}\} \quad (2)$$

where $E\{(\cdot)\}$ denotes the expectation operator. Since y_n is determined from e_{n-1} and y_{n-1} by the predictor, it can be omitted and (2) can be rewritten also as:

$$\hat{x}_n = E\{x_n|e_n,e_{n-1},y_{n-1}\} \quad (3)$$

Since the prediction and the estimation functions are implicit (as we shall see later), it is proposed in [1,2] to store them in look-up tables (LUT's). This way there is no computational load at all but only memory accesses. Consequently, a very fast system can be implemented with quite modest memory requirements, and can handle high sampling rates as required, for example, in video processing.

The performance of the above scheme for image compression is described in [1] and the results for speech signals are given in [2], both in comparison with DPCM (non-adaptive), for 1 bit/sample and 2 bits/sample. No adaptive version of the above scheme is proposed in [1,2].

4. Prediction and Estimation Characteristics

In this section we describe in detail the prediction and estimation characteristics for a 1 bit/sample system. In the sequel we consider also the 2 bits/sample system.

Let $\{x_n\}$ be a zero mean Gaussian process with variance σ^2 and correlation coefficient $\rho \triangleq E(x_n x_{n+1})/\sigma^2$, and assume that these parameters are known.

4.1 The predictor

In order to ensure equal probabilities for the quantization levels at the transmitter, the predictor: $y_{n+1}=M(y_n,e_n)$ must be the median of the conditional density function $g(x_{n+1}|x_n \geq y_n)$ for $e_n=1$, or the median of the density function $g(x_{n+1}|x_n < y_n)$ for $e_n=0$. For the case $e_n=1$, it follows from the Bayes formula that:

$$g(x_{n+1}|x_n \geq y_n)=\frac{g(x_{n+1},x_n \geq y_n)}{P(x_n \geq y_n)} \quad (4)$$

The denominator of (4) is given by $Q(\frac{y_n}{\sigma})$, where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-t^2/2}\,dt \,. \quad (5)$$

For the numerator of (4) we have:

$$g(x_{n+1},x_n \geq y_n)=\int_{y_n}^{\infty} g(x_{n+1},\vartheta)\,d\vartheta =$$

$$= g(x_{n+1})\int_{y_n}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma}\exp\left\{-\frac{(\vartheta-\rho x_{n+1})^2}{2\sigma^2(1-\rho^2)}\right\}d\vartheta =$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x_{n+1}^2}{2\sigma^2}\right)Q\left(\frac{y_n-\rho x_{n+1}}{\sigma\sqrt{1-\rho^2}}\right) \quad (6)$$

By putting (6) into (4) and integrating with respect to x_{n+1} from y_{n+1} to infinity, we obtain an expression for $Pr\{e_{n+1}=1|e_n,y_n\}$. Now, by (1) we have the following equation:

$$\Pr\{x_{n+1} \geq y_{n+1}|x_n \geq y_n\} =$$

$$\left[\sqrt{2\pi}\sigma Q\left(\frac{y_n}{\sigma}\right)\right]^{-1}\int_{y_{n+1}}^{\infty} e^{-\vartheta^2/2\sigma^2}Q\left(\frac{y_n-\rho\vartheta}{\sigma\sqrt{1-\rho^2}}\right)d\vartheta=\frac{1}{2} \quad (7)$$

In expression (7) the variables ρ,σ and y_n can be viewed as parameters and y_{n+1} can be viewed as the unknown. This equation can be solved by numeric techniques. For the case $e_n=0$ a similar equation is obtained by using symmetry considerations. The resulting predictor is non-linear as is demonstrated in Fig.2.

4.2 The Estimator

The estimator reconstructs the input by assigning the appropriate representation levels for the predictor's quantization decision levels. These levels are the centroids [6] of the ranges of the input samples given the information $\{e_n,y_{n-1},e_{n-1}\}$. For the case $e_n=e_{n-1}=1$ it can be shown (using integration by parts) that (3) satisfies:

$$\hat{x}_n=\sigma\sqrt{\frac{2}{\pi}}\frac{1}{Q\left(\frac{y_{n-1}}{\sigma}\right)}\left\{e^{-y_n^2/2\sigma^2}Q\left[\frac{y_{n-1}-\rho y_n}{\sigma\sqrt{1-\rho^2}}\right]+\right.$$

$$\left.+\rho\, e^{-y_{n-1}^2/2\sigma^2}Q\left[\frac{y_n-\rho y_{n-1}}{\sigma\sqrt{1-\rho^2}}\right]\right\} \quad (8)$$

For other values of e_n and e_{n-1}, similar expressions are obtained by using symmetry properties of the Gaussian distribution.

Because of practical limitations explained below, a 2 bits/sample system is not obtained just as a simple extension of the 1 bit/sample system, since there are now 3 threshold values. Here, if each threshold is represented by 8 bits, and the error is quantized to 2 bits, then the predictor LUT is addressed by 26 bits. Consequently, the predictor LUT size needed is 192 Mbyte ! Clearly, one must limit the number of states. This could be done by dividing the support of the density functions $g(x_{n+1}|x_n \geq y_n)$ and $g(x_{n+1}|x_n < y_n)$ into four non-overlapping intervals, each having a probability of 1/4. This solution is of course suboptimal.

5. System Performance for Stationary processes

In this section we present some simulation results performed to measure the ME performance for stationary Gaussian processes at the rates of 1 and 2 bits/sample.

We now examine the dependence of the SNR on the value of the correlation coefficient - ρ. The SNR is defined as

$$\mathrm{SNR} \triangleq 10\log_{10}\left[\left(\sum_{n=1}^{Nk}x_n^2\right)/\left(\sum_{n=1}^{Nk}(x_n-\hat{x}_n)^2\right)\right] \quad (9)$$

where N denotes the number of points per sequence and k - the number of sequences. The values of ρ used were $\rho=0.2, 0.5, 0.8, 0.9, 0.95, 0.98$. For each of these values, $k=50$ Gauss-Markov sequences were produced. Each sequence had $N=4096$ points. In addition, for each value of ρ the prediction and estimation LUT's were computed. The upper bound for the SNR of the reconstruction of a 1st order Gauss-Markov process from any representation by R bits/sample can be easily obtained from the rate-distortion function [3].

$$\mathrm{SNR}[dB] \leq 6.02R - 10\log_{10}(1-\rho^2) \quad (10)$$

In Fig.3 several graphs of SNR vs. ρ are presented. The figure compares the performances of ME, DPCM and the upper bound provided by (10) for the above rates.

The stepsizes for DPCM and LDM (Linear Delta Modulation) were selected empirically to minimize the MSE [5].

Several inferences are drawn from these curves:

1. The performances of all the systems examined are considerably far from the upper bound.
2. For a wide range of ρ, the ME system significantly overperforms DPCM and LDM, particularly at the rate of 1 bit/sample.

It is seen from Fig.3 that for a wide range of the correlation coefficient (zero to 0.8 or 0.85) the ME system has a higher SNR, by up to 5dB than LDM at 1 bit/sample, and about 3 dB above DPCM for 2 bits/sample.

6. The Adaptive Scheme

In order to make the system described earlier adaptive, one needs to estimate ρ and σ at each time instant, and to update the predictor and the estimator, accordingly, at both transmitter and receiver.

6.1 Gain Adaptation

It is suggested to estimate the parameter σ as follows:

$$S_n = \lambda S_{n-1} + (1-\lambda)\tilde{x}_n^2 \qquad (0 \le \lambda < 1) \qquad (11)$$

where: $\hat{\sigma}_n = \sqrt{S_n}$ is the estimate of σ at time n, $\tilde{x}_n$ - the reconstructed "normalized" signal (see Fig.4) and λ - a decay which determines the speed of adaptation.

In this way, the variance of the "normalized" signal remains roughly constant. At the receiver, $\tilde{x}_n$ is multiplied by σ_n ($\hat{x}_n = \tilde{x}_n \cdot \sigma_n$). The main problem observed in using this AGC mechanism, is a limit cycle effect which occurs when the input signal has a low energy. The limit cycle is characterized by high amplitude oscillations in the reconstructed waveform. Since the gain is adapted using the *reconstructed* signal, it turns out that the gain (σ_n) does not decay sufficiently in low energy intervals, and these oscillations remain large. To overcome the limit cycle problem it is first necessary to identify this event and to force the variable S_n to decrease, regardless of the value of the reconstructed signal. We have therefore modified (11) as follows:

$$S_n = \begin{cases} \lambda_1 S_{n-1} + (1-\lambda_1)\tilde{x}_n^2 & \text{, no limit cycle exists} \\ \lambda_2 S_{n-1} & \text{, limit cycle exists} \end{cases} \qquad (12)$$

where λ_1 and λ_2 are positive numbers smaller than 1. The identification of a limit cycle occurrence is based on the alternating sign of the prediction variable y_n. A limit cycle event is declared whenever the sign of y_n alternates at least three times successively. This mechanism was found to remove completely the limit cycle effect.

6.2 Adaptation of the Correlation Coefficient (ρ)

To adapt the system to variations in the correlation coefficient ρ, it is suggested to estimate ρ in the following way:

$$C_n = \lambda C_{n-1} + \tilde{x}_n \tilde{x}_{n-1} \qquad (13a)$$

$$S_n = \lambda S_{n-1} + \tilde{x}_n^2 \qquad (13b)$$

$$\hat{\rho}_n = C_n / S_n \qquad (13c)$$

where λ is the "forgetting" factor ($0 < \lambda \le 1$). It is necessary to limit the values of ρ_n such that $|\hat{\rho}_n|$ would not exceed unity.

The main problem now is how to use this estimate ($\hat{\rho}_n$) to update the predictor and the estimator. We have seen (expression (8)) that the estimate $\hat{x}_n$ can be expressed "explicitly" in terms of ρ, σ, y_n, e_n, e_{n-1} and y_{n-1}, where y_n is related to y_{n-1} and e_{n-1} by the prediction function. However, for the predictor we do not have an explicit formula. Therefore, it is proposed to use a simple approximation of the predictor by an explicit function. This is easily done by defining the predictor as the conditional expected value (instead of the median value), e.g., for $e_n = 1$, we have:

$$y_{n+1} \cong E(x_{n+1} | x_n \ge y_n) = \frac{\rho\sigma}{\sqrt{2\pi}} \frac{\exp\left(-\frac{y_n^2}{2\sigma^2}\right)}{Q\left(\frac{y_n}{\sigma}\right)} \qquad (14)$$

This approximation turns out to be a very good one (particularly for large values of $|y_n|$). Simulation results did not reveal any significant differences between using the exact predictor or the above approximated predictor. Similar ideas can be used for the 2 bits/sample system [7]. We now have "explicit" formulas for both the prediction and estimation by which updated values of ρ and σ can easily be substituted.

The computational load required to adapt the system is heavier than in classical ADPCM because these formulas are quite complicated. A lookup table for the function $Q(\cdot)$ is needed as well. But, as it was shown above, if the input is not too highly correlated, that adaptive version of the ME method overperforms ADPCM for non-stationary Gaussian processes with slowly varying covariance functions. For speech signals, the adaptive ME scheme turns out to perform equivalently to ADPCM in terms of quality and intelligibility.

References

[1] E. Angel and L. Daigle, "A High Speed Maximum Entropy Encoder for Images", IEEE ICASSP, 1983, pp. 1236-1239.

[2] E. Angel, L. Daigle and M. Rodriguez, "A Maximum Entropy Encoder for Speech", IEEE ICASSP, 1983, pp. 1292-1295.

[3] T. Berger, "Rate Distortion Theory", Prentice-Hall, Cliffs N.J., 1971.

[4] N.S. Jayant and P. Noll, "Digital Coding of Waveforms", Englewood Cliffs, N.J., Prentice-Hall, 1984.

[5] L.R. Rabiner and R.W. Schaffer, "Digital Speech processing", Prentice-Hall Inc. Englewood, Cliffs, N.J., 1978.

[6] Y. Linde, A. Buzo and R.M. Gray, "An Algorithm for Vector Quantization Design", IEEE Trans. on Communication, Vol. COM-28, No.1, Jan. 80, pp. 84-95.

[7] N. Merhav, Adaptive Maximum Entropy Coding of Speech Signals, M.Sc. Dissertation, Technion - I.I.T., Haifa, Israel, Nov. 1985. (In Hebrew).

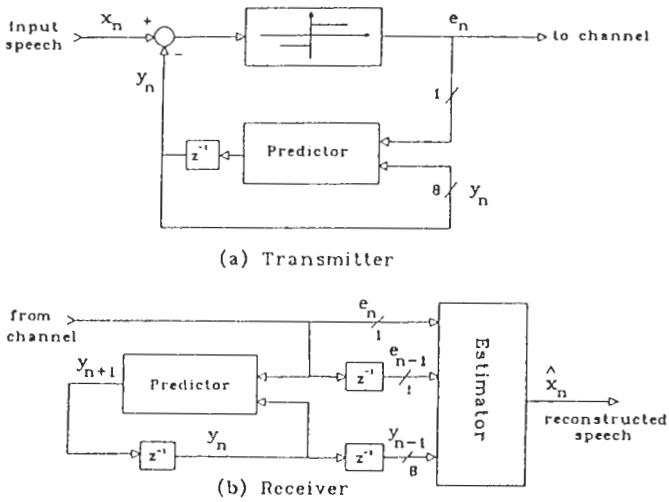

Fig. 1: - The compression scheme proposed in [1,2].

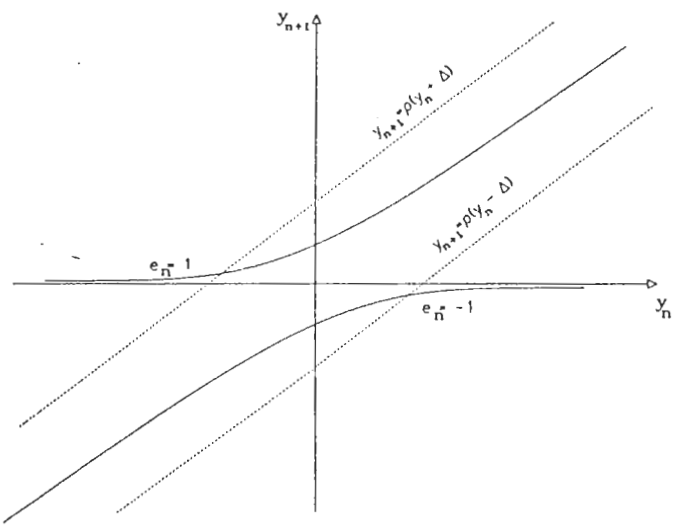

Fig. 2: - Prediction characteristics of ME and LDM (Linear Delta Modulator) for $\rho = 0.9$: solid line - ME predictor, dashed line - LDM predictor

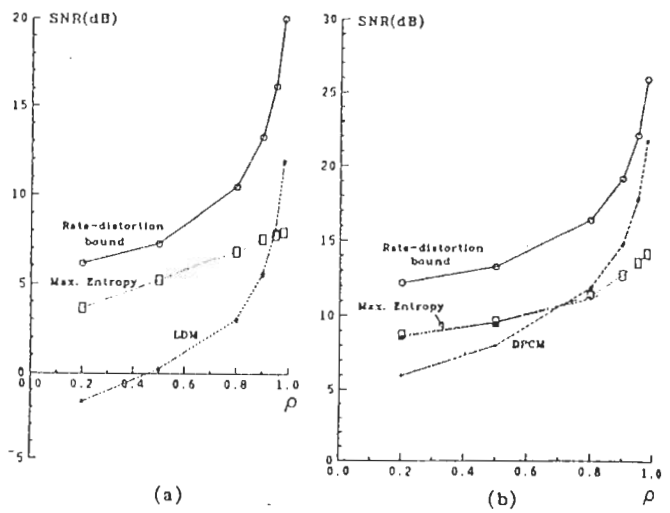

Fig. 3: - SNR versus ρ for the various systems: (a) R = 1 bit/sample (b) R = 2 bits/sample

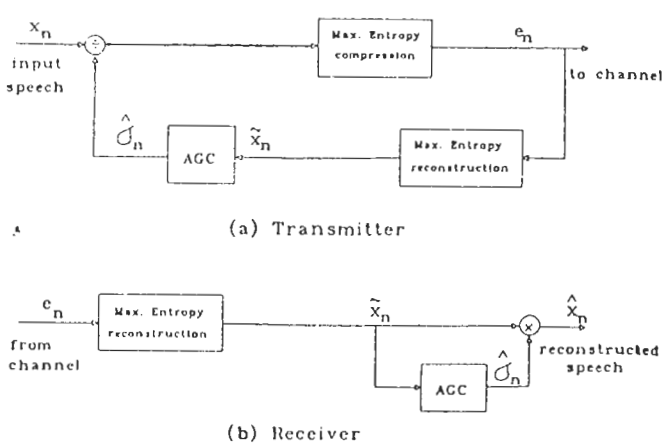

Fig. 4: - Transmitter and receiver with AGC.