ON THE NONLINEARITY OF LINEAR PREDICTION

Gernot Kubin
Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology, Gusshausstrasse 25/389, A–1040 Vienna, Austria
E-mail: g.kubin@ieee.org

ABSTRACT

This paper analyzes adaptive linear prediction and the effects of the underlying optimality criterion on the prediction error. It is well known that the signal-dependent optimization process converts the linear filter into a nonlinear signal processing device and that this will influence the statistics of the filter output in a way not expected from linear filter theory. For minimum-phase L_p-optimal linear predictors, we can show that the prediction error is maximally close to an i.i.d. process whose probability density function is given by $A\exp(-\lambda|x|^p)$. This result is applied to linear predictive analysis-by-synthesis coding of speech and to predictive decision-feedback equalization of channels with nongaussian noise. Implications for testing time series for linearity or gaussianity are discussed, too.

1. INTRODUCTION

Linear prediction is widely used in communications systems. In source coding, it maps a given waveform onto a residual signal with less correlation among the signal samples. This simplifies quantization of the waveform and reduces the quantization noise variance by the *prediction gain.* In channel equalization, a *noise predictor* helps to minimize the noise gain of decision-feedback equalizers and to accommodate delayed symbol decisions in the same structure. On a more general level, linear prediction is the standard tool for fitting linear autoregressive (AR) models to time series.

The implementation of linear predictors is usually based on linear FIR filters whose coefficients are automatically obtained from adaptation algorithms. This results in a dependence of the filter characteristics on some signal properties and shows that *the overall adaptive filter operates as a nonlinear system.*

It is common practice to abstract from this nonlinearity and to analyze adaptive filters in terms of linear filter theory. The direct effects of signal-dependent adaptation on the processed signals themselves are rarely considered. It is the purpose of this paper to address these effects and to demonstrate how the optimality criterion underlying the adaptation process determines the statistical properties of the prediction error.

2. THEORY

We develop our theoretical analysis in four steps. Details of proofs are omitted here for brevity.

Lemma 1 ([1]) *Any L_2-optimal linear one-step prediction-error filter*

$$e(n) = x(n) - \sum_{k=1}^{N} a_k x(n-k) \tag{1}$$

is minimum phase *(including the case $N \to \infty$). As shown in [1], the minimum-phase property can be preserved for a wider class of weighted frequency-domain cost functions, too.*

Lemma 2 ([2]) *A stationary stochastic process $\{X(n)\}$ with a constrained L_p norm (or p-th absolute moment) of its continuous first-order probability density f_X,*

$$L_p(X) = \mathrm{E}\{|X|^p\} = \int f_X(x)|x|^p \,\mathrm{d}x = \textit{constant}, \tag{2}$$

achieves the maximum differential entropy rate $\bar{h}(X)$ *iff it is an i.i.d. process with*

$$f_X(x) = A\exp(-\lambda|x|^p). \tag{3}$$

Corollary 1 ([3]) *Conversely, if the differential entropy rate $\bar{h}(X)$ of a stationary stochastic process $\{X(n)\}$ is constrained to a given constant value, then only the i.i.d. process with first-order density f_X as given in eq. (3) achieves the* minimum L_p norm.

Lemma 3 *If a stationary stochastic process $\{X(n)\}$ is* filtered by a minimum-phase system *with impulse response h_k, $k = 0, 1, 2, \ldots$ and $h_0 = 1$, the differential entropy rate $\bar{h}(Y)$ of its output is identical to the input rate $\bar{h}(X)$:*

$$\bar{h}(Y) = \bar{h}(X) + \log h_0 = \bar{h}(X). \tag{4}$$

In particular, this identity holds for the entropy rates of the original signal $\{X(n)\}$ and the prediction error $\{E(n)\}$.

Definition 1 *The* relative entropy rate *(or Kullback Leibler distance) $\bar{D}(X||Y)$ between two stationary stochastic processes $\{X(n)\}$ and $\{Y(n)\}$ is defined by [4]*

$$\bar{D}(X||Y) = \lim_{M\to\infty} \frac{1}{M} \int f_{\mathbf{X}(n)}(\mathbf{z}) \log \frac{f_{\mathbf{X}(n)}(\mathbf{z})}{f_{\mathbf{Y}(n)}(\mathbf{z})} \,\mathrm{d}\mathbf{z}, \tag{5}$$

where $f_{\mathbf{X}(n)}$ and $f_{\mathbf{Y}(n)}$ are the M-dimensional joint density functions of the vectors $\mathbf{X}(n) = [X(n), X(n-1), \ldots, X(n-M+1)]^{\mathrm{T}}$ and $\mathbf{Y}(n) = [Y(n), Y(n-1), \ldots, Y(n-M+1)]^{\mathrm{T}}$, respectively.

It can be shown that the relative entropy rate is always nonnegative, $\bar{D}(X||Y) \geq 0$, and that $\bar{D}(X||Y) = 0$ if the processes $\{X(n)\}$ and $\{Y(n)\}$ have the same probability structure. For gaussian processes, relative entropy rate equals the Itakura-Saito distortion between their power spectra [4, 5].

Theorem 1 *For a stationary process $\{X(n)\}$, the error process $\{E(n)\}$ of a minimum-phase linear predictor eq. (1) under the L_p criterion is* maximally close *to an i.i.d. process $\{Z(n)\}$ with first-order density as in eq. (3), in other words, their relative entropy rate $\bar{D}(E||Z)$ is minimum:*

$$\bar{D}(E||Z) = \underbrace{-\bar{h}(E)}_{-\bar{h}(X)} + \underbrace{\lambda \int f_E(z)|z|^p \,\mathrm{d}z}_{L_p(E)} - \log A = \min. \tag{6}$$

The relative entropy rate $\bar{D}(E||Z)$ of eq. (6) can be decomposed into two additive, nonnegative components:

$$\bar{D}(E||Z) = D(f_E||f_Z) + \bar{R}(E), \tag{7}$$

where the relative entropy $D(f_E||f_Z)$ measures the distance between the first-order densities of $\{E(n)\}$ and $\{Z(n)\}$ and where the redundancy rate

$$\bar{R}(E) = h(E) - \bar{h}(E) = \tag{8}$$

$$= \lim_{M\to\infty} I(E(n); E(n-1), \ldots, E(n-M)) \tag{9}$$

measures the mutual information left among the prediction error samples.

Note that the redundancy rate $\bar{R}(E)$ is a nonnegative quantity even for continuous-valued signals as considered here. From the decomposition of the relative entropy rate, the optimum linear predictor may trade between the two objectives of matching the first-order density of the error signal to a target density and of removing the statistical dependencies of the original waveform.

3. APPLICATIONS

3.1. Synthetic signals

Apparently, a positive relative entropy rate $\bar{D}(E||Z) > 0$ is always an indicator of some *mismatch* between the observed process and the adaptive system (characterized by three features, i.e., the *linear structure* of the predictor, its *order*, and the optimization *criterion*). We will start with the perfectly matched situation and proceed to various mismatched situations.

Linear L_p-matched AR processes are defined as linear autoregressive processes with an i.i.d. innovation process that has a density of the type eq. (3). If the predictor order N is not less than the order of the linear AR system generating the process the relative entropy rate $\bar{D}(E||Z)$ vanishes for a cost function L_p which is the negative log-likelihood of the innovation process density. As a special (deterministic) case, a sum of K sinusoids with $2K \leq N$ leads to a vanishing prediction error for any L_p criterion.

Gaussian processes constitute a remarkable signal class as their *amplitude distribution is not changed by linear filtering.* Therefore, the distance of the densities $D(f_E||f_Z)$ cannot be influenced by the filter optimization. The second component is simple to evaluate for gaussian processes where the redundancy rate is given by

$$\bar{R}(E) = -0.5 \log \Xi(E) \tag{10}$$

with the spectral flatness $\Xi(E)$ as defined in [5, eq. (6.24)]. In this case, the filter adaptation will maximize the spectral flatness of the residual no matter which p is chosen for the L_p norm.

Nongaussian i.i.d. processes have a *zero redundancy rate* $\bar{R}(X) = 0$, so one expects that an optimum predictor leaves us with an error process identical to the predicted process. However, this is not true in general if the first-order density of the predicted process has certain asymmetries. The simplest example is the L_2-optimal first-order prediction of a noncentral i.i.d. process where the optimum predictor coefficient a_1 is given by

$$a_1 = \frac{\mathrm{E}^2[X]}{\mathrm{E}[X^2]}. \tag{11}$$

From this, the redundancy rate of the error $\bar{R}(E)$ *increases* to a positive value *(i.e., adaptive linear prediction introduces statistical dependencies into the prediction error which are absent from the original process!)* whereas the distance $D(f_E||f_Z)$ *decreases* at the same time such that their sum $\bar{D}(E||Z)$ is still minimized. Here, a better match to the (gaussian) target density is achieved at the price of more redundancy in the error.

Nondeterministic processes which are neither gaussian nor i.i.d. will typically neither achieve a match in the marginal density ($D(f_E||f_Z) > 0$) nor in the dependence structure ($\bar{R}(E) > 0$). In this situation, the filter adaptation process may attain a smaller relative entropy rate $\bar{D}(E||Z)$ by reducing either the distance $D(f_E||f_Z)$ or the redundancy rate $\bar{R}(E)$ or both.

Deterministic processes such as countable sums of sinusoids or chaotic signals are characterized by an *infinite redundancy rate* $\bar{R}(X) = \infty$. Unless the predictor can model the signal exactly, the linear prediction error remains a nonzero deterministic signal with infinite redundancy rate $\bar{R}(E) = \infty$. In this *undermodeling* case, the minimization of an L_p norm of the error can only be achieved by minimizing the distance of the densities $D(f_E||f_Z)$. Therefore, the dependency of the error statistics on the optimization criterion will be most pronounced for this case. As an example, the choice of a least-mean-squares criterion (LMS or L_2 norm) will result in an (almost) gaussian prediction error density whereas a least-mean-absolute criterion (LMA or L_1 norm) will result in an (almost) Laplace density.

3.2. Testing for gaussianity/linearity

The statistical properties of the prediction error may largely depend on the optimization criterion used for predictor design. This effect sheds doubt on the usual practice to indirectly test the gaussianity or the (linear/nonlinear) dependence structure of a time series on its linear modeling residual. While it is true that only a gaussian process will lead to a prediction error which is exactly a gaussian process, application of the popular L_2 norm for predictor design leaves us always with an error that is maximally close to a white gaussian process. Note that this result goes far beyond the usual observation [6] that linear filtering of an i.i.d. process increases its gaussianity: *out of all linear, minimum-phase filters of a certain order N, the least-squares prediction error filter maximizes the 'white-gaussianity' of its output no matter what the input statistics are.*

From this and the discussion after eq. (11), testing procedures such as the Hinich test [7] may actually run into numerical difficulties when applied to the linear modeling residual rather than to the original process. Our theoretical analysis thus explains some experimental observations reported in the chaos theory literature [8]. As a general caveat, one should always use a non-L_2 norm for predictor optimization when running a conventional test for gaussianity on its residual.

3.3. Source coding of speech

Linear least-squares prediction is the core of *linear predictive analysis-by-synthesis (LPAS)* speech coders [9] which are widely used in digital cellular telephony. Already the first studies [10, 11] reported that the linear prediction residual has a first-order density which apparently is gaussian. This led to two conclusions: first, that the speech waveform itself may have gaussian statistics and second, that the codebook used to model the prediction residual can be populated with gaussian random numbers and still achieve a good waveform match. While the second observation is correct the first observation is an overgeneralization as seen from voiced speech

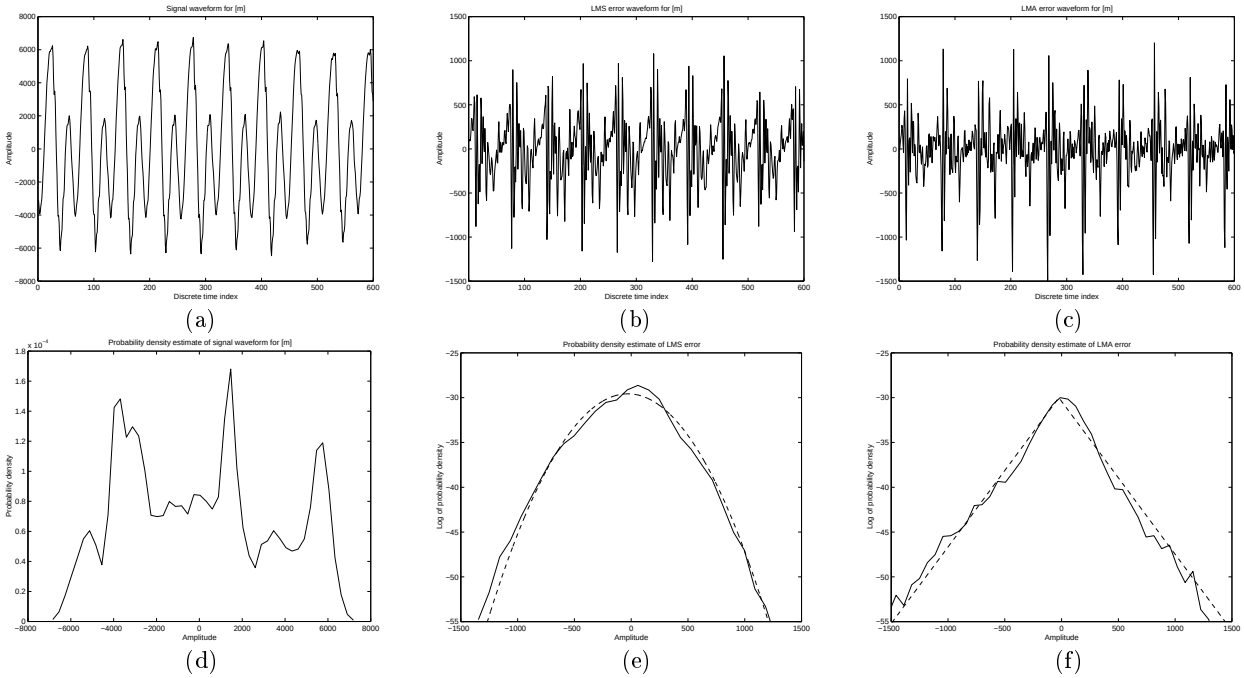

Figure 1. Time-domain waveforms and histogram estimates of amplitude probability densities for linear prediction of sustained sound [m], sampled at $f_s = 8$ kHz, total duration 2 seconds, predictor order $N = 10$. (a) Original waveform, (b) LMS error waveform, (c) LMA error waveform (d) original signal histogram (linear scale), (e) LMS error histogram (log scale, solid = estimated, dashed = theoretical), and (f) LMA error histogram (log scale, solid = estimated, dashed = theoretical). Waveforms show only a central 75 ms segment of the data whereas histograms use all 16,000 data samples.

sounds. Their first-order density is usually multimodal, cf. fig. 1.

Plot (a) shows a time-domain waveform segment of the sustained consonant [m] and plot (d) its histogram (based on 16,000 samples) which clearly shows three distinct modes, i.e., a significant deviation from a gaussian distribution. The other two signals are prediction residuals obtained with 10th order linear predictors which differ only in their optimization criteria: the LMS criterion (L_2) in plots (b) and (e) and the LMA criterion (L_1) in plots (c) and (f). Filter optimization was performed with recursive adaptation algorithms, the *unnormalized LMS* algorithm for the L_2 norm and the *signed error* algorithm for the L_1 norm [12]. The striking result is that the residual histograms closely match their theoretical minimum L_p densities plotted with dashed lines. Therefore, no conclusions about the 'true' innovation density should be made in this case. Rather, voiced speech can be approximated by a sum of sinusoids and the chosen order N is far too low to represent all the harmonics of this speech sound. From section 3.1, a strong dependence of the error density on the cost function is expected here. This dependence can be exploited in speech coder design where a desired amplitude distribution of the prediction residual can be achieved via proper cost function selection.

3.4. Decision feedback equalization

Decision feedback equalization (DFE) can be viewed as an enhancement to a forward linear equalizer that allows to minimize the noise variance at the input to the decision device while maintaining the same residual intersymbol interference (ISI). It can also be interpreted as cascading a linear equalizer with a linear prediction error filter [13, 14]. While a DFE implementation need not follow this interpretation, it becomes advantageous if significant delay of the decision device (e.g., due to Viterbi decoding of trellis-coded modulation signals) has to be accommodated. A signal flow diagram is given in fig. 2 which also defines the forward adaptation error $e_1(n)$, its prediction $\hat{e}_1(n)$, and the prediction error $e_2(n)$.

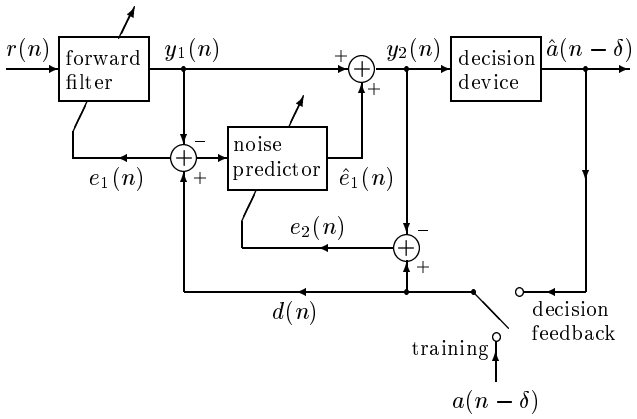

Figure 2. Predictive DFE structure: received signal $r(n)$, equalizer outputs $y_1(n) + \hat{e}_1(n) = y_2(n)$, adaptation errors $e_1(n), e_2(n)$, symbol sequences $a(n), \hat{a}(n)$.

We have simulated a single-carrier binary phase-shift keying system with an FIR equivalent baseband channel model (an order 5 lowpass) and additive deterministic interference (a rectangular pulse with normalized frequency $\theta_{\mathrm{IF}} = 0.1$) at an SNR of 5 dB. The forward filter has $M = 40$ coefficients, the noise predictor $N = 5$ coefficients. For comparison purposes, the forward filter is always adapted with the same algorithm, i.e., recursive least squares with growing window (forgetting factor $\lambda = 1$) and we use the training mode (at a delay $\delta = 22$) throughout the simulation.

For the noise predictor, the influence of the optimality criterion on the statistics of the residual noise becomes significant as the usual mean-square error (MSE) is not the ultimate performance measure. Most of the probability mass of the residual $e_2(n)$ should be concentrated within the receiver's decision thresholds to minimize the error probability.

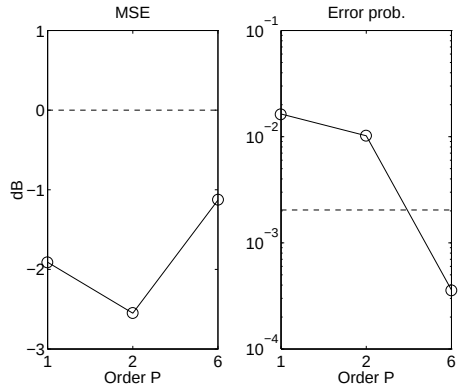

Figure 3. Mean-square error (MSE) of prediction residual and symbol error probability of predictive DFE after convergence for various optimality criteria L_p with $p = 1, 2, 6$. The circles o indicate results obtained by averaging over several simulation runs of 20,000 samples each where the last 10,000 samples have been used for steady-state measurements. The dashed line indicates the performance of the linear forward equalizer alone (without decision feedback).

Therefore, the preferred optimization criterion for the linear predictor is an L_p norm with $p \gg 1$.

Figure 3 shows the steady-state performance results for three choices $p = 1, 2, 6$ using gradient-type algorithms as described in [12, 15]. It is clearly seen that the L_2 optimal adaptation algorithm minimizes the MSE after the noise predictor while the L_6 algorithm has the highest MSE but its symbol error probability is reduced by more than an order of magnitude. The improvement for the L_6 algorithm is so pronounced because the interference is deterministic and the order $M = 5$ predictor cannot cancel it completely. With gaussian noise, the same performance as with the L_2 norm is preserved. In that sense, *the choice of a high L_p norm increases the robustness of the adaptive DFE.*

4. DISCUSSION

We have presented an information-theory based analysis of how the choice of a cost function for optimum linear prediction influences the statistics of the prediction error. For the most prominent special case, we have proved that least-squares linear prediction will always yield a prediction residual which is maximally close to a white gaussian process. Along this line, we have used a measure for the distance between the probability structures of two stationary stochastic processes, relative entropy rate, which generalizes spectral flatness theory from the L_2 case to more general optimization criteria.

This observation opens the question whether there exists *a best cost function* that should be recommended for general use, i.e., when there is no prior knowledge of the process statistics. Such questions have received previous consideration under the title of *minimum-entropy deconvolution* [6], *projection pursuit* [16], or *independent component analysis* [17]. One answer is the *minimum entropy cost function.* Minimization of the first-order differential entropy of the prediction error for a given entropy rate is directly equivalent to minimization of the error's redundancy rate. Unlike conventional L_p optimization, this optimization approach circumvents the mismatch between the cost function and the process statistics. For instance, a minimum-entropy linear predictor will never introduce extra statistical dependencies into the prediction residual. It is, therefore, the method of choice if tests for the probability structure of the given process are applied to the prediction residual.

On a more general level, we have demonstrated that information theory is useful for analyzing both linear and nonlinear signal processing algorithms like the signed error algorithm. This analysis is robust as it implicitly includes the nongaussian/nonlinear case. Thus we can predict the system behavior for situations where the adaptive linear system is not matched to the process structure.

REFERENCES

[1] W. F. G. Mecklenbräuker, "Remarks on the minimum phase property of optimal prediction error filters and some related questions," *IEEE Signal Process. Lett.*, submitted 1997.

[2] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* Tokyo etc.: McGraw-Hill, 2nd ed., 1984.

[3] G. Kubin, "Nonlinear autoregressive modeling based on rate-distortion theory," in *Adaptive Methods and Emergent Techniques for Signal Processing and Communications* (D. Docampo and A. Figueiras, eds.), (Vigo, Spain), pp. 391–398, June 1993.

[4] M. Pinsker, *Information and Information Stability of Random Variables and Processes.* Moscow: Ize. Akad. Nauk. SSSR, 1960. English translation: San Francisco, CA: Holden-Day, 1964.

[5] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech.* Berlin, Heidelberg, New York: Springer, 1976.

[6] D. Donoho, "On mimimum entropy deconvolution," in *Applied Time Series Analysis II* (D. Findley, ed.), pp. 565–608, Academic Press Inc., 1981.

[7] R. A. Ashley, D. M. Patterson, and M. J. Hinich, "A diagnostic test for nonlinear serial dependence in time series fitting errors," *J. Time Series Anal.*, vol. 7, no. 3, pp. 165–178, 1986.

[8] J. Theiler and S. Eubank, "Don't bleach chaotic data," *Chaos*, vol. 3, pp. 771–782, 1993.

[9] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 70–119, Amsterdam (The Netherlands): Elsevier, 1995.

[10] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Comm.*, vol. COM-30, pp. 600–614, 1982.

[11] H. Reininger and D. Wolf, "Speech and speaker independent codebook design in VQ coding schemes," in *Proc. ICASSP'85*, (Tampa, FL), pp. 1700–1702, Mar. 1985.

[12] A. Gersho, "Adaptive filtering with binary reinforcement," *IEEE Transactions on Information Theory*, vol. IT-30, pp. 191–199, March 1984.

[13] D. G. Messerschmitt, "A geometric theory of intersymbol interference, part I: Zero-forcing and decision-feedback equalization," *B.S.T.J.*, vol. 52, pp. 1483–1519, Nov. 1973.

[14] J. G. Proakis, *Digital Communications.* New York, etc.: McGraw-Hill, 3rd ed., 1995.

[15] E. Walach and B. Widrow, "The least mean forth (LMF) adaptive algorithm and its family," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 275–283, Mar. 1984.

[16] P. J. Huber, "Projection pursuit," *Annals of Statistics*, vol. 13, pp. 435–525, 1985.

[17] P. Comon, "Independent component analysis: A new concept," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.