A DETERMINISTIC ANNEALING APPROACH TO SWITCHED PREDICTOR DESIGN FOR ADAPTIVE COMPRESSION SYSTEMS

Bharath Vishwanath, Tejaswi Nanjundaswamy and Kenneth Rose

Department of Electrical and Computer Engineering
University of California Santa Barbara, CA 93106
{bharathv, tejaswi, rose}@ece.ucsb.edu

ABSTRACT

Adaptive prediction is important in the compression of non-stationary signals, and a common remedy is to switch between appropriately designed prediction modes. This paper presents a near optimal procedure to design prediction modes for an adaptive compression system. The main challenges include: instability and mismatched statistics during closed loop design; and the severe non-convexity of the cost function trapping the system in poor local minima. The statistical mismatch is circumvented through a largely open loop (hence stable) design that is devised to asymptotically optimize the prediction modes for closed loop operation. The non-convexity of the cost function is handled by the deterministic annealing paradigm, a powerful non-convex optimization framework devised to avoid poor local minima. Experimental results provide substantial gains validating the efficacy of the proposed design technique.

Index Terms— predictor design, deterministic annealing, asymptotic closed-loop

1. INTRODUCTION

Linear Prediction is widely used in multimedia compression systems including speech and video coding [1–3]. Most signals of interest are non-stationary in nature. This motivates block or frame based encoding, where the input signal is partitioned into blocks, and optimal prediction filters are designed for each block. However, sending the actual prediction filters as side information to the decoder results in a severe rate penalty. Thus, a standard practice is to define a fixed selection of prediction filters (modes) at the encoder and only send an index informing the decoder of the selected mode. This is effectively "quantization" of the filter space to produce a small codebook of representative filters. This scenario is common, for example, in video coding and its spatial and temporal prediction. There is, therefore, strong motivation for an efficient procedure for offline design of prediction modes.

Mode switching at the encoder is a non-linear operation, which makes for a challenging non-convex optimization problem, as the derivative of the cost function with respect to mode decisions vanishes almost everywhere. Thus, the predictor design is done in an iterative manner, similar to the "K-means" clustering algorithm [4], a flavor of which can be seen in an earlier work from our lab in [5]. In an iteration, the input blocks are assigned to prediction modes that minimize the prediction residual. The codebook is then refined by designing optimal prediction filters for each cluster of input blocks. Since the predictors are applied to reconstructed samples at the decoder, the design of prediction modes depends on the reconstructed samples, which themselves depend on the prediction, thus giving a first glimpse of the closed loop conundrum we discuss below. In principle, upon design of prediction modes, the reconstructed samples must be updated and the process must be repeated until (and if) it converges. Clearly, the prediction loop creates complex relation between the prediction filters and the reconstructed samples and makes the design more challenging. In the context of joint design of predictors and quantizers, two early approaches have been proposed in [6] and [7]. In open loop design, the actual input signal is used for predictor design. However, since the decoder does not have access to original input samples, during operation the predictor must operate on the reconstructed samples and is hence mismatched to its input statistics. In closed loop design, the predictor is designed for the reconstructed samples of the previous iteration, but it is then applied to the different reconstructed samples resulting from the new predictor, leading to mismatch that can catastrophically grow due to the prediction loop, which represents a major stability problem. To address this, asymptotic closed loop (ACL) design was proposed in [8]. ACL operates in open loop fashion by predicting from reconstructed samples in previous iteration, exactly the statistics based on which the predictor was updated. However, on convergence, the reconstructed samples remain unchanged across iterations, effectively optimizing the system for closed loop operation. Although ACL resolves the design stability issues, it still suffers from the non-convexity of the cost surface and often converges to poor local minima.

This paper presents a deterministic annealing (DA) approach to the design of prediction modes. Inspired by princi-

ples of statistical physics and information theory, DA was proposed as a powerful non-convex optimization framework [9]. The probabilistic nature of DA yields an effective cost function via expectation, which is differentiable with respect to the prediction modes. Its annealing schedule gradually reduces the randomness of the solution so as to avoid poor local minima. The overall proposed method embeds ACL within the DA framework. The careful annealing schedule of DA is complemented by the stable design platform of ACL, effectively addressing the central design challenges enumerated above. Experimental results provide evidence for substantial performance gains.

2. BACKGROUND

2.1. Prediction model

Fig. 1 shows a predictive compression system. Let x_n, $0 \leq n \leq N$ be the input samples. The signal is modelled as first-order auto-regressive process. Current sample x_n is predicted from the previous reconstructed sample as,

$$\tilde{x}_n = \alpha \hat{x}_{n-1} \tag{1}$$

The resulting prediction error, i.e, $x_n - \tilde{x}_n$ is quantized and sent to the decoder. The sum of squared prediction errors that needs to be minimized is given by,

$$E = \sum_{n=1}^{N} (x_n - \alpha \hat{x}_{n-1})^2 \tag{2}$$

The optimal predictor is given by,

$$\alpha = \frac{\sum_n x_n \hat{x}_{n-1}}{\sum_n \hat{x}_{n-1}^2} \tag{3}$$

In order to adapt the predictor to the variations in the signal statistics, let the input be partitioned into blocks (frames) Let N_f be the set of samples belonging to a particular frame f. Let the encoder be given a choice of K prediction filters $\{\alpha_k\}, k = 1, 2..K$. The encoder chooses the best prediction mode for each block of samples. Let the best prediction mode for a given block f be $\hat{\alpha}_f$. The problem at hand is to design the prediction filters $\{\alpha_k\}$ such that the overall mean squared prediction error is minimized i.e,

$$E = \sum_f \sum_{n \in N_f} (x_n - \hat{\alpha}_f \hat{x}_{n-1})^2 \tag{4}$$

2.2. Iterative K-mode predictor design

Let us assume for the moment that we have a set of reconstructed samples $\hat{x}_n$ at the encoder. Given these reconstructions, we can design prediction modes in a way similar to "K-means" clustering. With an initialization of the prediction modes, the following steps are performed iteratively:

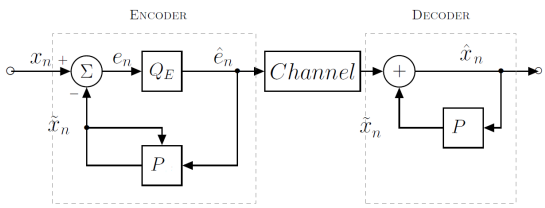

Fig. 1. Predictive compression system

- Mode assignment: For a given block f, assign the best mode from the set of prediction modes which minimizes the squared prediction error for the block.
- Prediction modes update: Let N_k be the union of samples from frames that share the same prediction mode. Similar to (3), the optimal prediction mode α_k for this cluster is given by,

$$\alpha_k = \frac{\sum_{n \in N_k} x_n \hat{x}_{n-1}}{\sum_n \hat{x}_{n-1}^2} \tag{5}$$

With the new set of prediction modes, the reconstructed samples at the encoder are updated. These steps are repeated until convergence. The reconstructions can be updated in different ways, leading to the following design paradigms.

2.3. Open-loop, closed-loop and asymptotic closed-loop design

Various techniques have been proposed in the context of joint design of predictors and quantizers. Since in most of modern codecs the quantizer is fixed (up to scaling), our focus here is on predictor design given fixed quantizers. In open loop predictor design (see e.g., [6]), the predictor is designed using original samples. However, since the predictor must be applied to reconstructed samples, to avoid decoder drift, it is in fact operating on statistics mismatched with the design. In closed loop design [7], predictors are designed iteratively. Let $\hat{\alpha}_f^{i-1}$ be the predictor for frame f in iteration $i-1$. The reconstructed samples for the corresponding frame in iteration i is updated as,

$$\hat{x}_n^i = \hat{\alpha}_f^{i-1} \hat{x}_{n-1}^i + \hat{e}_n^i \tag{6}$$

where $\hat{e}_n^i$ is the quantized prediction error $e_n = x_n - \hat{\alpha}_f^{i-1} \hat{x}_{n-1}^i$. Predictor $\hat{\alpha}_f^{i-1}$ was designed for reconstruction in iteration $i-1$: $\{\hat{x}_n^{i-1}\}$. However, it is applied to the reconstructed samples of iteration i: $\{\hat{x}_n^i\}$. This mismatch results in design instability, which grows with feedback through the prediction loop and often proves catastrophic at low rates. To tackle this issue, ACL was proposed in [8]. ACL enjoys the best of both worlds. At each iteration, the samples are predicted and reconstructed in open loop fashion as,

$$\hat{x}_n^i = \hat{\alpha}_f^{i-1} \hat{x}_{n-1}^{i-1} + \hat{e}_n^i \tag{7}$$

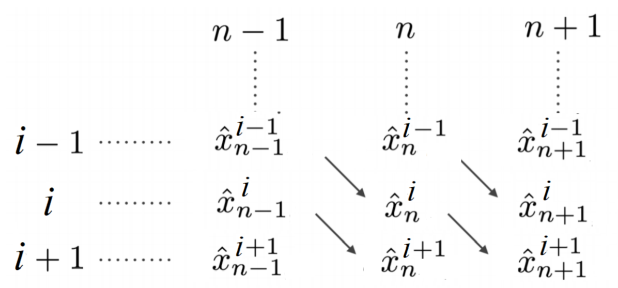

Fig. 2. Asymptotic closed loop design

where $\hat{e}_n^i$ is the quantized prediction error $e_n = x_n - \hat{\alpha}_f^{i-1}\hat{x}_{n-1}^{i-1}$. The predictor $\hat{\alpha}_f^{i-1}$ is used with reconstructed samples $\hat{x}_n^{i-1}$, the same set of samples that it was designed for, thereby providing a stable design platform. The new set of reconstructed samples are then used to design prediction modes α_k^i. Upon convergence, the reconstructed samples remain the same over iterations. Thus predicting from $\hat{x}_n^{i-1}$ is same as predicting from $\hat{x}_n^i$, which is essentially closed loop operation. Fig. 2 illustrates ACL design.

3. PROPOSED METHOD

The hard prediction mode assignment in (4) makes it difficult to optimize the system with respect to prediction modes, since the derivative vanishes almost everywhere. Hence an iterative K-mode design was originally proposed. However, this only ensures convergence to a local minimum and renders the system highly sensitive to initialization. A related problem is encountered in quantizer design, where the piecewise linear nature of the quantizer makes it a challenging optimization problem. In this paper, we propose to embed the ACL based minimization of the overall prediction error within the DA framework, in order to jointly overcome all the above fundamental design challenges. The proposed approach is inspired by, and builds on the DA framework of [9]. DA is motivated by the intuition gained from annealing processes in physical chemistry, where certain systems are driven to their low energy states by gradual cooling of the system. Analogously, we introduce controlled randomness in the prediction mode assignment for the blocks, but deterministically minimize the overall prediction error, thereby avoiding many poor local minima. The amount of randomness is measured by the Shannon entropy and is essentially controlled by the "temperature" of the system. The prediction mode assignment is no longer non-linear, and is differentiable everywhere paving the way to effective optimization of prediction modes.

We consider a random setting wherein in each frame, a prediction mode is chosen *in probability*. The mean squared prediction error to minimize in ACL iteration i is written as

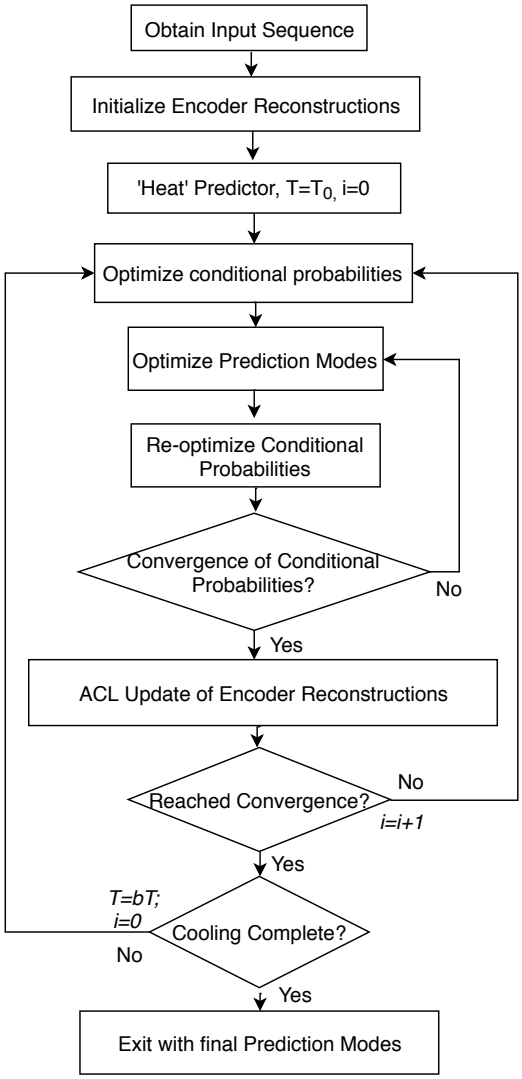

Fig. 3. Flow chart of the proposed algorithm

the expectation,

$$J = \sum_f \sum_k \sum_{n \in N_f} P_f P_{k|f}^i (x_n - \alpha_k^i \hat{x}_{n-1}^i)^2 \qquad (8)$$

where P_f is the probability of the input data frame and is assumed to be uniform. Association probability $P_{k|f}^i$ is the probability that prediction mode α_k is selected for input frame f. The degree of randomness in the system is measured by the Shannon entropy as,

$$H = -\sum_f \sum_k P_{fk}^i \log(P_{fk}^i), \qquad (9)$$

where $P_{fk}^i = P_f P_{k|f}^i$ is the joint distribution of prediction modes and training data blocks. The problem is viewed as the minimization of the Lagrangian cost function, directly analogous to the Helmholtz free energy of statistical physics:

$$F = J - TH \qquad (10)$$

The degree of randomness is controlled by Lagrangian parameter T, which is the temperature in the physical analogy. As we lower T, we trade entropy for prediction error. At the limit of zero randomness, we in fact directly minimize the overall prediction error. Minimizing the Lagrangian cost with respect to the association probabilities $P^i_{k|f}$, while additionally imposing the constraint $\sum_k P^i_{k|f} = 1$, yields the Gibbs distribution:

$$P^i_{k|f} = \frac{e^{\frac{-\sum_{n\in N_f}(x_n - \alpha^i_k \hat{x}^i_{n-1})^2}{T}}}{\sum_j e^{\frac{-\sum_{n\in N_f}(x_n - \alpha^i_j \hat{x}^i_{n-1})^2}{T}}} \tag{11}$$

Note that at high temperatures, we in fact maximize the system entropy and indeed the association probabilities are uniform.

The optimal prediction modes now satisfy,

$$\begin{aligned}\frac{\partial J}{\partial \alpha^i_k} &= \sum_f \sum_{n\in N_f} 2P_f P^i_{k|f}(x_n - \alpha^i_k \hat{x}^i_{n-1})(-\hat{x}^i_{n-1}) \\ &= 0\end{aligned} \tag{12}$$

Thus, the optimal prediction modes are given by,

$$\alpha^i_k = \frac{\sum_f \sum_{n\in N_f} P^i_{k|f} x_n \hat{x}^i_{n-1}}{\sum_f \sum_{n\in N_f} P^i_{k|f} (\hat{x}^i_{n-1})^2} \tag{13}$$

At high temperatures, given the uniform association probabilities, it follows from (13) that all the prediction modes are coincident and are all equal to the optimal single prediction mode of (3), *regardless of initialization*. As the temperature is lowered, the association probabilities become more "discriminating" and the solution more deterministic, with the emergence of different prediction modes through a mechanism of "phase transitions" in the physical analogy.

The reconstructed samples $\hat{x}^{i+1}_n$ in a block f are now updated in ACL fashion as,

$$\hat{x}^{i+1}_n = \sum_k P^i_{k|f}(\alpha^i_k \hat{x}^i_{n-1} + \hat{e}^i_{n,k}) \tag{14}$$

where, $\hat{e}^i_{n,k}$ is the quantized prediction error. The overall design procedure is summarized in Fig. 3.

4. SIMULATION RESULTS

The proposed method is applicable to any predictive compression system. For a proof of concept in a simple setting we focus on scalar first-order prediction and chose speech signals as source data. A set of six speech files from the EBU SQAM database were chosen for simulations [10]. Half of the speech files were used as training set for designing prediction modes and the remaining half as the test set. A set of six prediction modes were designed. A fixed dead-zone quantizer was employed for quantization. Different R-D points were obtained by varying Lagrange multiplier of entropy constrained quantization. The 3 competitors were: closed-loop (CL), "plain ACL", and the proposed method (DA-ACL). While DA-ACL is independent of initialization, CL and ACL designs were repeated with different initialization and the best results were selected. Fig. 4 shows the reconstructed SNR versus bit rate. It is evident from the results that the proposed DA-ACL method gives significant 0.4dB and 5dB gains over competitors ACL and CL, respectively.

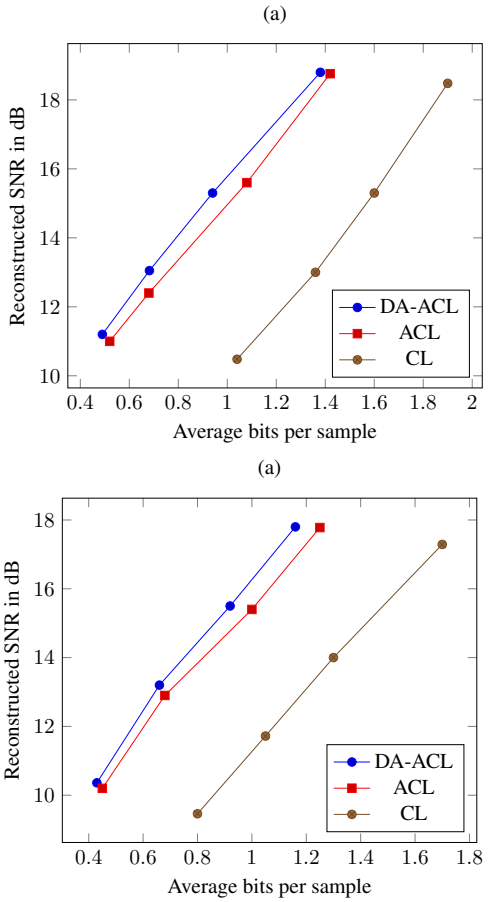

Fig. 4. Reconstructed SNR vs average bits per sample for (a) training set (b) test set

5. CONCLUSIONS

This paper presents a novel approach to optimal design of prediction modes for adaptive compression systems. It eliminates the major shortcomings of existing approaches: statistical mismatch, design instability, and poor local minima. Significant gains over both training and test sets across a range of bit rates substantiates the utility of the proposed approach.

6. REFERENCES

[1] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, et al., "Overview of the high efficiency video coding(hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[5] S. Li, Y. Chen, J. Han, T. Nanjundaswamy, and K. Rose, "Rate-distortion optimization and adaptation of intra prediction filter parameters," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3146–3150.

[6] V. Cuperman and A. Gersho, "Vector predictive coding of speech at 16 kbits/s," *IEEE Transactions on Communications*, vol. 33, no. 7, pp. 685–696, 1985.

[7] P-C. Chang and R. Gray, "Gradient algorithms for designing predictive vector quantizers," *IEEE transactions on acoustics, speech, and signal processing*, vol. 34, no. 4, pp. 679–690, 1986.

[8] H. Khalil, K. Rose, and S. L. Regunathan, "The asymptotic closed-loop approach to predictive vector quantizer design with application in video coding," *IEEE transactions on image processing*, vol. 10, no. 1, pp. 15–23, 2001.

[9] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[10] G.T. Waters, "Sound quality assessment material recordings for subjective tests," *users handbook for the ebu-sqam compact disc, European Broadcasting Union, Avenue Albert Lancaster*, vol. 32, pp. 1180.