NECESSARY CONDITIONS FOR OPTIMAL TRELLIS-CODED RESIDUAL VECTOR QUANTIZATION

Mohammad A. U. Khan

Department of Electrical Engineering
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
maukhan@kfupm.edu.sa

ABSTRACT

The union of residual vector quantization (RVQ) and trellis-coded vector quantization (TCVQ) was considered by various authors where the emphasis was on the sequential design. In this paper, we consider a new jointly optimized combination of RVQ and TCVQ with advantages in all categories. Necessary conditions for optimality of the jointly optimized trellis-coded residual vector quantization (TCRVQ) are derived. Simulation results for jointly optimized TCRVQ are presented for memoryless Gaussian, Laplacian and uniform sources. The rate-distortion performance is shown to be better than RVQ and sequentially designed TCRVQ.

1. INTRODUCTION

Vector quantization (VQ) has been studied extensively for data compression. However, direct use of VQ results in a heavy memory and complexity burden that is unattractive. These computation and memory demands required by VQ depend on the VQ output rate R, and the vector dimension n. Structurally unconstrained vector quantizers have exponentially dependent costs proportional to 2^{nR}.

Relief can be obtained by imposing carefully selected structural constraints. A simple and efficient type of structurally constrained VQ is Residual Vector Quantization (RVQ) or multistage VQ [1], which consists of a cascade of vector quantization (VQ) stages. The first stage input vector $\boldsymbol{x}^1$ is quantized to generate the approximation $\hat{\boldsymbol{x}}^1$. The difference vector is then computed to form the residual $\boldsymbol{x}^2 = \boldsymbol{x}^1 - \hat{\boldsymbol{x}^1}$, which serves as the input to the next stage, and so on. An RVQ with P stages and N_p vectors per stage can uniquely represent $\prod_{p=1}^{P} N_p$ vectors, which can amount to orders of magnitude less memory than a conventional VQ. Similar saving in computation may also be realized by exploiting the RVQ structure. However, the efficient structure impacts the quality of performance as one would expect [2], [3], the severity of which is directly related to the loss in degrees of freedom.

A jointly optimized fixed-rate RVQ design approach was presented in [4], in which the RVQ stages are optimized jointly. In that design, an attempt was made to minimize the overall quantization error of the RVQ in lieu of merely optimizing the individual stages in isolation. In [5], the theory of fixed-rate RVQ was extended to the case of variable-rate RVQ. The resulting RVQ design, known as entropy-constrained RVQ (EC-RVQ), was able to provide performance superior to that of unstructured, exhaustive search, entropy-constrained VQ for a given specification of memory and complexity. These performance gains come from the fact that implementation efficiencies of RVQ can be used to push the practical peak rates of the entropy constrained encoder to higher values [5].

It was shown in [6] that trellis-coded quantization (TCQ) has the ability to achieve better cell shapes by using scalar quantization along a trellis structure. In fact, according to [7], the performance is better than the lattices known to date up to 24 dimensions. The success of TCQ in achieving better cell shapes prompted researchers to consider exploring the combination of TCQ and residual vector quantization.

This combination was considered in [8] and [9], where the authors proposed residual trellis-coded vector quantization (RTCVQ) with a trellis in each residual stage. In these papers, encoding is performed sequentially, stage by stage, without regard to the overall system output. The design method adopted by the authors optimizes each residual stage in isolation, and consequently does not design an optimal trellis-based RVQ. The sub-optimality of the sequential design becomes more apparent when used in a large residual stage setup. We will refer to this scheme as sequentially designed trellis-coded residual vector quantization.

Trellis-coded residual vector quantizers were also presented and analyzed in [10], [11], where experimental results on natural sources were presented. This approach, although sequential in its design, is different in that a single trellis structure is used that extends along all the residual stages, providing stage symbols for each input vector. Since the input vectors are coded independently rather than using the trellis, this structure does not exploit any interaction among neighboring vectors.

In this paper, we develop the theory of a jointly optimized trellis-coded residual vector quantization (TCRVQ) that employs direct sum codebook and joint optimization over all residual stages. The first part of the paper (Section 2) introduces residual vector quantization, sequentially optimized trellis-coded residual vector quantization and jointly optimized trellis-coded residual vector quantization. The next part of the paper gives a derivation of necessary con-

dition for the optimal jointly optimized trellis-coded residual vector quantization. Section 3 reports comparisons and simulation results for Laplacian, Gaussian, and uniform sources.

2. SOME PRELIMINARIES

2.1. Residual vector quantizers

Residual vector quantization is associated with its direct sum codebook, which can be constructed by enumerating and summing over the tree paths embodied in the stage structure. The quantizer is specified by a triple $(A^e, Q^e, \mathcal{P}^e)$, consisting of a direct sum codebook, direct sum mapping, and direct sum partition, respectively. The elements of the direct sum codebook A^e are the elements of the set of all possible sums of stage code vectors i.e., $A^e = A^1 + A^2 + \cdots + A^P$, one code vector summed from each residual stage. The code vectors comprising direct sum codebook, $\boldsymbol{y}^e \in A^e$ are indexed by the P tuples $\boldsymbol{j}^P = (j^1, j^2, \ldots, j^P)$, and can be written as

$$\boldsymbol{y}^e(\boldsymbol{j}^P) = \sum_{p=1}^{P} \boldsymbol{y}^p(j^p), \tag{1}$$

where $\boldsymbol{y}^p(j^p)$ represents the j^pth code vector of the pth stage codebook. The direct sum partition $\mathcal{P}^e$ is the collection of all direct sum cells $S^e(\boldsymbol{j}^P)$. The union of direct sum cells covers the n-dimensional space and has the property that $S_j \cap S_k = \emptyset$ for $j \neq k$.

The direct sum mapping $Q^e : \mathcal{R}^n \longmapsto A^e$, replaces each vector input $\boldsymbol{x}^1$ with a direct sum codebook vector $\boldsymbol{y}^e(\boldsymbol{j}^P)$. The average distortion of the residual vector quantizer is

$$D(\boldsymbol{x}^1, \hat{\boldsymbol{x}}^1) = \int d[\boldsymbol{x}^1, Q^e(\boldsymbol{x}^1)] dF_{\boldsymbol{X}^1}, \tag{2}$$

where $\hat{\boldsymbol{x}}^1$ is the quantized representation of the input vector $\boldsymbol{x}^1$ and $F_{\boldsymbol{X}^1}$ is the source distribution function.

A necessary condition for minimum distortion is derived in [4], [12]. This condition implies that each stage code vector is obtained as a conditional mean of residual random vectors where residuals are formed from encoding decisions of both *prior* and *subsequent* stages. For a P-stage residual vector quantizer, the ρth stage residual ξ^ρ, also called the grafted residual, is defined as

$$\xi^\rho = \boldsymbol{x}^1 - \sum_{p=1, p\neq\rho}^{P} \boldsymbol{y}^p(j^p). \tag{3}$$

The necessary condition described above forms the basis for the design of jointly optimized trellis-coded residual vector quantization. The jointly optimized RVQ is an improvement over the sequentially optimized RVQ proposed earlier. It derives its stage code vectors from the centroids of residuals obtained through encoding decisions of only *prior* stages, as defined below

$$\boldsymbol{x}^\rho = \boldsymbol{x}^1 - \sum_{p=1}^{\rho-1} \boldsymbol{y}^p(j^p). \tag{4}$$

Based on the necessary condition, jointly optimized RVQ can be designed iteratively where each iteration consists of first optimizing the encoder while holding the decoder fixed, and then optimizing the decoder while holding the encoder fixed. Since each step of this procedure can only reduce or leave unchanged the average distortion of the encoder/decoder pair, the design process converges to a local minimum.

3. NECESSARY CONDITION FOR A TRELLIS-CODED DIRECT SUM CODEBOOK

A necessary condition for minimum distortion of a trellis-coded direct sum codebook can be described in the following way. P^{te} is held fixed by keeping the path maps in the trellis structure fixed. Once the path map (i.e the trellis search) is held fixed, the stage sub-codebooks are optimized jointly. We will describe the conditions for the scalar case and the mean-square error fidelity criterion. These conditions can be extended to the vector case in a straightforward manner.

Let X be a real random variable with probability density function $f_X(\cdot)$. We wish to find a locally optimal set of expanded codebooks $\{A^{tp}\}$ by finding $y^{K,p}(j^{K,p})$ that gives a locally minimum value of

$$D_{mse} = \int_{-\infty}^{\infty} [x - Q^{te}(x)]^2 f_x(x) dx, \tag{5}$$

based on a trellis search.

The trellis-coded direct sum code vectors available depend on which sub-codebook is labeled along that branch. Thus, each trellis state will provide a particular subset of codebooks on its outgoing branches, constituting a residual quantizer. In other words, we can say that after partitioning two RVQs are constructed. Depending on which trellis state we are in, a particular RVQ will result. The trellis-coded direct sum sub-codebooks constituting each RVQ should be optimized jointly. Alternatively, the above equation can be written more precisely as

$$\begin{aligned} D_{mse} &= \sum_{\boldsymbol{j}^{1,P}} \int_{S^{te}(\boldsymbol{j}^{1,P})} [x - y^{te}(\boldsymbol{j}^{1,P})]^2 f_X(x) dx \\ &+ \sum_{\boldsymbol{j}^{2,P}} \int_{S^{te}(\boldsymbol{j}^{2,P})} [x - y^{te}(\boldsymbol{j}^{2,P})]^2 f_X(x) dx. \end{aligned} \tag{6}$$

In addition to assuming a fixed P^{te}, assume that all expanded stage codebooks except for $A^{t\rho}$ with $\rho \in \{1, \ldots, P\}$ are held fixed. To minimize D_{mse} with respect to the $g^{K,\rho}$th code vector in $A^{t\rho}$, set the partial derivative of both terms in equation (6) with respect to $y^{K,\rho}(g^{(K,\rho)})$ equal to zero. We will get two equations, one for each RVQ. For brevity, we will use a general setup with variable K to represent both RVQ with $K = 1$ for the first RVQ and $K = 2$ for the second. Thus, we have

$$\sum_{\boldsymbol{j}^{K,P}} \int_{S^{te}(\boldsymbol{j}^{K,P})} [x - y^{te}(\boldsymbol{j}^{K,P})][\frac{\partial y^{te}(\boldsymbol{j}^{K,P})}{\partial y^{K,\rho}(g^{(K,\rho)})}] f_X(x) dx = 0. \tag{7}$$

This partial derivative in brackets is

$$\frac{\partial y^{te}(\boldsymbol{j}^{K,P})}{\partial y^{K,\rho}(g^{(K,\rho)})} = \begin{cases} 1, & \text{if } \boldsymbol{j}^{K,P} \in H^{K,\rho}(g^{(K,\rho)}), \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $H^{K,\rho}(g^{(K,\rho)})$ is the set of all $\boldsymbol{j}^{K,P}$ such that the (K,ρ)th element of $\boldsymbol{j}^{K,P} = (j^{(K,1)}, j^{(K,2)}, \ldots, j^{(K,\rho)}, \ldots, j^{(K,P)})$ is equal to $g^{(K,\rho)}$. Solving for $y^{K,\rho}(g^{(K,\rho)})$ gives the desired result i.e. $y^{K,\rho}(g^{(K,\rho)}) =:$

$$\frac{\sum_{\boldsymbol{j}^{K,P} \in H^{K,\rho}(g^{(K,\rho)})} \int_{S^{te}} (\boldsymbol{j}^{K,P})(x - \sum_{p=1, p\neq\rho}^{P} y^{K,p}(j^{(K,p)})) f_x(x)dx}{\sum_{\boldsymbol{j}^{K,P} \in H^{K,\rho}(g^{(K,\rho)})} \int_{S^{te}(\boldsymbol{j}^{K,P})} f_X(x)dx}. \tag{9}$$

This equation dictates that the two direct sum quantizers available at the outgoing branches of the trellis have to be jointly optimized for all residual stages. For this purpose, *grafted residuals* introduced in [4] are used. Once the quantizers along each state of the trellis are optimized, the next optimality condition deals with searching the trellis. Here we use an optimal search like the Viterbi algorithm [13]. These two conditions will provide a locally optimal trellis-coded direct sum quantizer. Some preliminary results are described in next paragraph.

Experimental results derived from simulated quantizers are presented. The fidelity criterion used is the mean square error normalized by the source variance. Plots are presented that show rate vs *signal-to-noise ratio*. Signal-to-noise ratio is measured in dB and defined as

$$\text{SNR(dB)} = 10 \log_{10} \frac{\sum_{i=1}^{N} (x_i)^2}{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2} \tag{10}$$

The simulation results for this section are generated using training sets of 1,000,000 two-dimensional vectors. This training set size seems to be reasonable as we found almost no difference in performance results run inside and outside the training set. A 16-state trellis is used with a residual structure searched by employing M=2 multi-path searching.

Table 1: Performance(SNR in dB) of various vector quantizers for the two-dimensional Laplacian source at 0.5, 1.0, 1.5, 2.0 bits/sample

Rate	RVQ	Seq. TCRVQ	TCRVQ	R(D)
0.5	1.60	2.47	2.47	N/A
1.0	3.59	4.21	4.99	6.62
1.5	5.91	6.18	7.42	N/A
2.0	7.49	8.12	9.94	12.66

Table 1 shows the performance for memoryless Laplacian sources. TCRVQ performs better than RVQ and sequential TCRVQ at all rates. Here each residual stage is contributing 0.5 bit/sample. We see that at two stages, the performance of TCRVQ is 0.7 dB better than sequential TCRVQ. At four stages, the difference is 1.8 dB. The difference in performance of TCRVQ and RVQ at four stages is 2.4 dB.

Table 2 shows a memoryless Gaussian vector source.

Table 2: Performance(SNR in dB) of various vector quantizers for the two-dimensional Gaussian source at 0.5, 1.0, 1.5, 2.0 bits/sample

Rate	RVQ	Seq. TCRVQ	TCRVQ	R(D)
0.5	1.67	2.63	2.63	3.01
1.0	4.40	4.95	5.10	6.02
1.5	6.92	7.38	7.83	9.00
2.0	9.43	9.72	10.49	12.04

Here we also observe the improved performance of TCRVQ as compared to sequential TCRVQ and RVQ. The performance gap between TCRVQ and sequential TCRVQ is about 0.15 dB at two stages and is 0.8 dB for four stages. This performance gap for Gaussian sources is much less that of Laplacian sources. Similarly the performance of TCRVQ is about 1 dB better than RVQ at four stages. The difference for Gaussian and Laplacian sources may be attributed to the difference of density shape advantage associated with two sources. Two-dimensional Gaussian sources provide density shape advantage of 1.14 dB while the advantage for a two-dimensional Laplacian source is 2.27 dB, which is almost twice as high as that for the Gaussian source.

Table 3: Performance (SNR in dB) of various vector quantizers for the two-dimensional Uniform source at 0.5, 1.0, 1.5, 2.0 bits/sample

Rate	RVQ	Seq. TCRVQ	TCRVQ	R(D)
0.5	2.04	2.89	2.89	N/A
1.0	6.03	5.48	6.29	6.79
1.5	8.10	7.92	9.56	N/A
2.0	12.04	10.23	12.41	13.21

Table 3 shows the performance for uniform vector sources. This source is used for experiments because it does not provide any density shape advantage. Uniform sources only provide cell shape advantage. Table shows a very small gap of about 0.5 dB between the rate-distortion point and TCRVQ at 0.5 bits per sample. This reflects the ability of the TCRVQ coder to capture a larger part of the cell shape advantage from the trellis structure. The performance of TCRVQ is again better than RVQ and sequential TCRVQ and much closer to the R(D) points. Sequential TCRVQ simply does not perform well above 1.5 bits per sample and its performance is lower than even RVQ.

4. REFERENCES

[1] B. Juang and A. Gray, "Multiple stage vector quantization for speech coding," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 597–600, April 1982.

[2] R. Baker, *vector quantization of digital images.* PhD thesis, Standford university, Standford, CA, 1984.

[3] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proceedings of IEEE*, vol. 73, pp. 1551–1581, Nov 1985.

[4] C. F. Barnes and R. L. Frost, "Vector quantizers with direct sum codebooks," *IEEE transactions on information theory*, vol. 39, pp. 565–580, March 1993.

[5] F. Kossentini, M. Smith, and C. Barnes, "Necessary conditions for the optimality of variable-rate residual vector quantizers," *IEEE trans. on information theory*, vol. 41, pp. 1903–1914, November 1995.

[6] M. Marcellin and T. Fischer, "Trellis coded quantization of memoryless and gauss-markov sources," *IEEE transactions on communications*, vol. 38, pp. 82–93, Jan 1990.

[7] M. W. Marcellin, *Trellis coded quantization: an efficient technique for data compression.* PhD thesis, Texas A & M university, December 1987.

[8] A. Aksu and M. Salehi, "Multistage trellis coded quantization (ms-tcq) design and performance," *IEE Proceeding of Communications*, vol. 144, pp. 61–64, April 1997.

[9] A. Aksu and M. Salehi, "Design, performance and complexity analysis of residual trellis-coded vector quantizers," *IEEE transactions on communications*, vol. 46, pp. 1020–1026, August 1998.

[10] G. Motta and B. Carpentieri, "A new trellis vector residual quantizer: Applications to image coding," in *Proceedings of ICASSP-97*, pp. 2929–2931, 1997.

[11] G. Motta and B. Carpentieri, "Trellis vector residual quantization," in *International conf. on signal processing applications and tecgnology (ICSPAT 97)*, (San Diego (CA)), September 1997.

[12] C. Barnes, *Residual Quantizers.* PhD thesis, Brigham Young University, Prov UT, Dec 1989.

[13] G. Forney-Jr., "The viterbi algorithm," *Proc. IEEE (invited paper)*, vol. 61, pp. 268–278, March 1973.