Variable-Dimension Vector Quantization

Amitava Das, Ajit V. Rao, and Allen Gersho

Abstract— **In many signal compression applications, the evolution of the signal over time can be represented by a sequence of random vectors with varying dimensionality. Frequently, the generation of such *variable-dimension* vectors can be modeled as a random sampling of another signal vector with a large but fixed dimension. Efficient quantization of these *variable-dimension* vectors is a challenging task and a critical issue in speech coding algorithms based on harmonic spectral modeling. We introduce a simple and effective formulation of the problem and present a novel technique, called *variable-dimension vector quantization* (VDVQ), where the input variable-dimension vector is directly quantized with a single *universal* codebook. The application of VDVQ to low bit-rate speech coding demonstrates significant gain in subjective quality as well as in rate-distortion performance over prior indirect methods.**

I. Introduction

VECTOR quantization (VQ) [1] is a well-known technique for encoding a *fixed-dimension* random vector. However, in many signal compression applications the evolution of the signal over time is represented by a sequence of *variable-dimension* random vectors. Efficient encoding of these variable-dimension vectors is a challenging task that is essential for effective design of state-of-the-art speech coders such as the *multiband excitation* (MBE) coder [2] and the *sinusoidal transform coder* (STC) [3]. In the MBE, for example, the spectral information is represented by a set of spectral magnitude samples taken at the harmonics of the estimated fundamental frequency, or pitch, F_o. As the pitch varies from frame to frame, the number of samples in this set, or the dimension of this *spectral shape vector* (SSV) varies. The subjective quality of such spectral coders almost entirely depends on how well these SSV's are quantized. An efficient compression scheme for variable-dimension vectors will significantly improve the performance of these speech coders.

Theoretically, the optimal VQ-based solution is *multicodebook VQ* (MCVQ), where a separate codebook is used for each possible dimension. (See [4] for a treatment of MCVQ). However, a practical implementation of MCVQ may not be feasible due to the large requirements of codebook memory and training. Most prior solutions to variable-dimension quantization [5]–[7] are variations of an *indirect* procedure that we generically call *dimension conversion VQ* (DCVQ), where the vector $\mathbf{S}$ with variable dimension L is first mapped into a fixed-dimension vector $\mathbf{Y}$ prior to applying VQ. An L-dimension estimate $\hat{\mathbf{S}}$ is obtained from the decoded vector $\hat{\mathbf{Y}}$ by a reverse transformation. The problem of DCVQ is that the overall distortion has two components: i) D_q, due to quantization and ii) D_c, due to dimension conversion. At low rates, this additional distortion, D_c, is an unwelcome burden.

We present a novel technique, called *variable-dimension vector quantization* (VDVQ) which, without any dimension-conversion, offers a direct and efficient solution to the problem of encoding variable-dimension vectors. VDVQ uses a single universal codebook and a training set of reasonable size (as compared to a large number of codebooks and an excessively large training set needed in MCVQ). VDVQ does not have the additional, dimension-conversion distortion, D_c, of DCVQ. Therefore, as the bit rate is increased, the VDVQ distortion approaches zero, whereas the DCVQ distortion approaches D_c. As we will see later, VDVQ outperforms prior indirect solutions, such as the complex combination of scalar quantization and VQ used in improved multiband excitation coding (IMBE) [2], and linear prediction (LP) modeling [5], a DCVQ method, in terms of rate-distortion performance.

II. Variable-Dimension Vector Quantization

We formulate the variable-dimension vector quantization problem as follows. Let $\mathbf{S}$ be the variable-dimension vector, formed from a randomly selected subset of the components of an underlying K dimension random vector, $\mathbf{X}$. Both the choice of indexes and the size, L, of this subset are random. The elements of $\mathbf{S}$ are the selected components of $\mathbf{X}$ in corresponding order. This random selection g($\mathbf{X}$) can be represented by a K dimension binary *selector vector* $\mathbf{Q}$, whose nonzero components are pointers to the selected components of $\mathbf{X}$. For example, if $K = 4$, $\mathbf{X} = (x_1, x_2, x_3, x_4)$ and $\mathbf{Q} = (0, 1, 0, 1)$, then $\mathbf{S} = g(\mathbf{X}) = (x_2, x_4)$.

A. VDVQ Encoding

Guided by the selector vector $\mathbf{Q}$, the variable-dimension vector $\mathbf{S}$ is mapped into the K-dimension space to form an *extended vector* $\mathbf{Z}$. For example, if $\mathbf{Q} = (0, 1, 0, 1)$ and $\mathbf{S} = (q, r)$, then $\mathbf{Z} = (0, q, 0, r)$. Then $\mathbf{Z}$ is compared to each codevector $\mathbf{Y}_j$ of the universal codebook to find the best match that minimizes the distortion, as follows:

$$d(\mathbf{Z}, \mathbf{Y}_j) = \sum_{k=1}^{K} Q[k] d_1(Z[k], Y_j[k]) \tag{1}$$

where $d_1(\cdot, \cdot)$ is a suitable distortion measure between two scalars. The index j^*, for which $d(\mathbf{Z}, \mathbf{Y}_j)$ is minimum over all $j = 1, 2, \cdots, N$, is selected.

Manuscript received August 15, 1995. This work was supported by Fujitsu Labs, Inc., the University of California MICRO program, Echo Speech Corporation, Rockwell International Corporation, Signal Technology, Inc. Speech Technology Laboratories, Texas Instruments, Inc., and Qualcomm, Inc. The associate editor coordinating the review of this letter and approving it for publication was Dr. V. Viswanathan.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: gersho@ece.ucsb.edu).

Publisher Item Identifier S 1070-9908(96)05531-9.

1070–9908/96$05.00 © 1996 IEEE

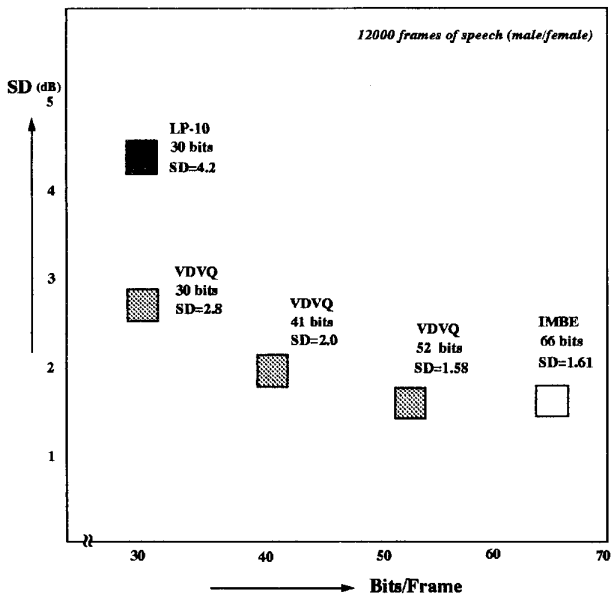

Fig. 1. VDVQ advantage: superior rate-distortion performance over existing methods.

B. VDVQ Decoding

The decoder receives the selector vector $\mathbf{Q}$ (or equivalently, F_o, in harmonic speech coding applications) and the optimal index j^* and computes an L-dimension estimate $\hat{\mathbf{S}}$ by subsampling $\mathbf{Y}_{j^*}$. Specifically, it selects the components of $\mathbf{Y}_{j^*}$ for which the corresponding components of $\mathbf{Q}$ are nonzero, proceeding in order of increasing index.

C. VDVQ Training

Given a large training set of pairs $\{(\mathbf{Q}_i, \mathbf{S}_i)\}$ and an initial codebook of size N and dimension K, the universal codebook is designed with an iterative procedure as in the generalized Lloyd algorithm (GLA) [1]. Let $\mathbf{Y}_j$, $j = 1, 2, \cdots, N$ be the codevectors prior to the current iteration. The two steps of each training iteration are as follows.

Nearest-Neighbor Clustering: i) For each training pair $(\mathbf{Q}_i, \mathbf{S}_i)$, form the extended vector $\mathbf{Z}_i$; ii) Assign $\mathbf{Z}_i$ to a cluster set C_m if $d(\mathbf{Z}_i, \mathbf{Y}_m) \leq d(\mathbf{Z}_i, \mathbf{Y}_j)$ for $j = 1, 2, \cdots, N$.

Centroid Computation: For each cluster$C_m, m = 1,2, \cdots, N$, find a new code vector, $\mathbf{Y}_m{}'$, the cluster centroid, that minimizes the cluster distortion $D_m = \sum_{\mathbf{Z}_j \in C_m} d(\mathbf{Z}_j, \mathbf{y})$ over all vectors, $\mathbf{y}$, in $\mathcal{R}^K$.

For the mean squared error distortion, where $d_1(s, y) = (s - y)^2$, the centroid rule gives

$$Y_m'[k] = \frac{\sum_{\mathbf{Z}_j \in C_m} \frac{1}{L_j} Q_j[k] Z_j[k]}{\sum_{\mathbf{Z}_j \in C_m} \frac{1}{L_j} Q_j[k]} \qquad \text{for} \qquad k = 1, 2, \cdots K \tag{2}$$

where L_j denotes the number of nonzero elements of $\mathbf{Q}_j$. A reasonably large training ratio (100 or so) is needed to ensure adequate representation for each component sample of the universal code vectors.

III. Performance of VDVQ in Speech Coding

In harmonic speech coders [2], let $\mathbf{W}$ represent the K magnitude values of the short-term spectrum of a frame of speech, where K is the number of discrete Fourier transform (DFT) points and $W[j]$ is a DFT magnitude sample. Then, let $\mathbf{U}$ be the vector of subsamples of $\mathbf{W}$ where each subsample is the nearest DFT sample to a pitch harmonic; and let $\mathbf{S}$ be the SSV defined in the log domain with $S[k] = 20\ log_{10} U[k], k = 1, 2, \cdots L$. Here, the DFT resolution determines the larger dimension K, whereas L, the variable dimension of $\mathbf{S}$ and the selector vector $\mathbf{Q}$ are completely determined by the fundamental frequency F_o (in Hz). Let F_s be the sampling frequency; then $\mathbf{Q}$ and $\mathbf{Z}$ are generated as

$$Q[k] = \begin{cases} 1 & \text{if } k = \lfloor 2jKF_o/F_s + 0.5 \rfloor \quad j > 0,\ 1 \leq k \leq K \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$Z[k] = \begin{cases} S[j] & \text{if } k = \lfloor 2jKF_o/F_s + 0.5 \rfloor\ j > 0,\ 1 \leq k \leq K \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

A *mean-removed split VQ* implementation of VDVQ [8] was applied to the 2400 b/s *enhanced multiband excitation* coder [9] and the 1400 b/s (average rate) *variable rate spectral coder*, [9]. In informal subjective testing [9], both coders attained speech quality comparable to or better than the 4150 b/s IMBE coder. Similar perceptual quality was observed with VDVQ at 30 b/frame, as opposed to 66 b/frame used in IMBE.

For objective comparison, we used the spectral distortion (SD) measure SD between the original SSV $\mathbf{U}$ and the quantized SSV $\hat{\mathbf{U}}$, defined by

$$SD = \left[\frac{1}{f_2 - f_1} \sum_{k=f_1}^{f_2} (20\ log_{10} U[k] - 20\ log_{10} \hat{U}[k])^2 \right]^{1/2} \tag{5}$$

in dB units, where (f_1, f_2) is the frequency range of interest normalized to number of samples of the DFT. (We cover the range 62.5–3562 Hz). Note that (5) is consistent with the usual definition of spectral distortion, and since the extended vector $\mathbf{Z}$ and the codevectors of the universal codebook are in the log-spectral domain, the VDVQ design, based on mean-square error, minimizes the SD.

Fig. 1 shows the SD of VDVQ at different bit rates (b/frame) and the SD of the IMBE quantization method and the tenth-order LP-method (LP-10) [5], a DCVQ method. The test corpus used in this comparison had 12 000 frames of male and female speech, outside of the training set. The results show that VDVQ outperforms LP-10 by 1.4 dB. VDVQ also outperforms IMBE in terms of rate-distortion measure, producing an SD of 1.58 dB at 52 b/frame as opposed to the 1.61 dB SD of IMBE at 66 b/frame. This performance gain is more remarkable since IMBE employs differential coding across frames, whereas the VDVQ scheme reported here does not.

IV. Conclusions

Variable-dimension vector quantization is a novel technique that offers a direct and efficient solution to the problem

of encoding variable-dimension vectors, outperforming prior indirect methods. The application of VDVQ to harmonic coding of speech demonstrates its effectiveness by achieving a significant gain in subjective quality as well as in rate-distortion performance. By integrating various forms of *structured vector quantization*, such as *tree-structured VQ*, *multi-stage VQ*, and *predictive VQ* [1], VDVQ can be customized to fit the performance objectives and memory and complexity constraints of a particular application. Such extensions are currently being explored.

REFERENCES

[1] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer, 1992.

[2] *Inmarsat-M Voice Codec Specifications*, Digital Voice Systems, Inc., Tech. Rep., ver. 2, 1991.

[3] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, 1986.

[4] J-P. Adoul and M. Delprat, "Design algorithm for variable-length vector quantizers," in *Proc. Allerton Conf. Circuits, Syst., Comput.*, 1986, pp. 1004–1011.

[5] M. S. Brandstein, "A 1.5 kbps multi-band excitation speech coder," M.S. thesis, MIT, Cambridge, MA, 1990.

[6] P. Lupini and V. Cuperman, "Vector quantization of harmonic magnitudes for low rate speech coders," in *Proc. IEEE Globecom Conf.*, 1994, vol. 2, pp. 858–862.

[7] M. Nishiguchi, J. Mutsumoto, R. Wakatsuki, and S. Ono, "Vector quantized MBE with simplified V/UV decision at 3.0 kbps," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, pp. 151–154.

[8] A. Das, A. Rao, and A. Gersho, "Variable dimension vector quantization of speech spectral for low-rate vocoders," in *Proc. IEEE Data Compress. Conf.*, 1994, pp. 420–429.

[9] A. Das and A. Gersho, "Variable dimensional spectral coding of speech at 2.4 kb/s and below with phonetic classification," in *Proc. IEEE Int. Conf. Acoust., Speech., Signal Processing*, 1995, pp. 492–495.