1990 International Symposium on Information
Theory and Its Applications (ISITA'90)
Hawaii, U.S.A., November 27-30, 1990

Robust Quantization of Memoryless Sources

Ashok C. Popat
Laboratoire de Traitement des Signaux, Department d'Electricité
Ecole Polytechnique Fédérale de Lausanne, CH-1015, SWITZERLAND

Kenneth Zeger
Dept. of Electrical Engineering
University of Hawaii, Honolulu, HI 96822

Abstract

A novel approach to quantizing discrete-time memoryless sources is presented. The method involves changing the amplitude distribution of the source to be approximately Gaussian by all-pass filtering, so that the source can be quantized (using a Lloyd-Max quantizer) more effectively than had it not been filtered. The filtered quantized source is passed through an inverse all-pass filter, so that the overall resulting quantization error is less than would be obtained by direct Lloyd-Max quantization of the source. An important feature is that the resulting performance is largely insensitive to errors in modeling the input PDF. The cost of this approach is some delay due to filtering.

1. Introduction

Much has been written about quantization of memoryless sources, in particular, Laplacian and gamma sources [1-4]. The subject is important because these sources are often used as models in image and speech coding [5][6]. An irony associated with the quantization of Laplacian and gamma sources is made evident by the graphs in Figure 1. Although the rate-distortion functions* of these sources are quite promising relative to, say, that of a Gaussian source, simple quantization** does not fulfill that promise. In fact, it can be seen that for any given rate, the Lloyd-Max quantizer achieves lower mean-square error for Gaussian sources than for Laplacian or gamma sources. This observation leads to quantization scheme described in this paper.

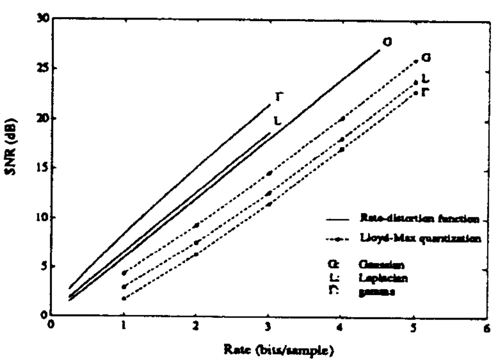

Figure 1: Performance (mean-square error versus rate) of Lloyd-Max quantization of Laplacian, gamma, and Gaussian sources, relative to the respective rate-distortion functions. Samples of the rate-distortion function for Laplacian and gamma sources were computed by means of the Blahut algorithm [10]. Quantizer performance figures are taken from [5, p. 135].

The present suggestion is to use simple quantization, but to filter the source before and after quantizing. If the filters are appropriately designed (see Section 5), then the filtered input signal will have an approximately Gaussian distribution, and the resulting quantization error of the overall system will approximate that for direct quantization of a Gaussian source. Moreover, since the initial filtering will tend to make *any* memoryless source appear Gaussian, the performance of the system is insensitive to errors in modeling the input. This robustness to the source statistics is a valuable feature, not normally present in quantization systems. Throughout this paper, it is assumed that the source is stationary and memoryless. For simplicity of notation, it is further assumed that the source has zero-mean and unit-variance.

2. Prior work

Many sophisticated alternatives to simple quantization have been suggested for memoryless sources. These alternatives include vector quantization [4], entropy-coded (variable-rate) quantization [2], and trellis coding [3][11]. While these techniques generally achieve better signal-to-noise ratios than the proposed scheme, they are also more complex, and do not offer the same degree of robustness.

The method of quantization proposed here appears to be novel, despite its stark simplicity. The only similar proposal of which the author is aware is one by Strube [12]. In that scheme, an all-pass filter is used in a speech ADPCM system to disperse pitch pulses over time, so that quantizer overload-distortion is reduced. Strube's scheme does *not* make use of an inverse filter, however, so that it does not result in a signal that approximates the original. The use of an all-pass prefilter and inverse postfilter to reversibly change the PDF of a signal has been suggested by Zenith [13], in the context of transmission of high-definition television signals. However, their suggestion has nothing to do with quantization.

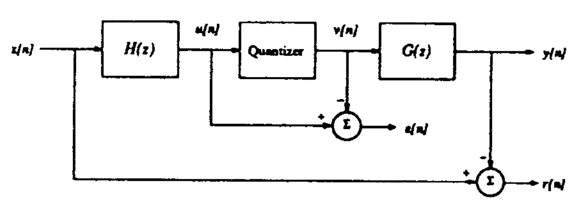

Figure 2: Quantization system employing a prefilter and a postfilter.

3. Preservation of quantization error

In this section it is shown that the mean-square error between input and output of the proposed system is nearly equal to that incurred by quantizing the signal after prefiltering.

Let $H(z)$ and $G(z)$ denote the z-transforms of the prefilter and postfilter, respectively, as shown in Figure 2, and let $h[n]$ and $g[n]$ denote the corresponding impulse responses. Let $r[n] = x[n-D] - y[n]$ denote the error of the overall system, where D is the delay due to filtering, and let $e[n]$ denote the error incurred by quantizing the

* For a given source, the rate-distortion function specifies the best possible performance attainable by any coding method [7, Chapter 9].
** Throughout, the term *simple quantization* refers to fixed-rate minimum-mean-square-error zero-memory quantization, or Lloyd-Max quantization [8][9].

1990 International Symposium on Information Theory and Its Applications (ISITA'90)
Hawaii, U.S.A., November 27-30, 1990

prefiltered signal $u[n]$ into $v[n]$. In the following analysis, it is *not* assumed that $e[n]$ is independent of $u[n]$.

By linearity of the post-filter,

$$\begin{aligned} y[n] &= g[n] * (u[n] - e[n]) \\ &= g[n] * u[n] - g[n] * e[n] \\ &= g[n] * h[n] * x[n] - g[n] * e[n], \end{aligned} \tag{1}$$

where $*$ indicates convolution. By hypothesis, $H(z)$ and $G(z)$ are approximate inverses of one another within a delay of D, so that (1) becomes

$$y[n] \approx x[n-D] - g[n] * e[n],$$

so that the error $r[n]$ for the overall system can be approximated as

$$r[n] \approx g[n] * e[n]. \tag{2}$$

By assumption, $x[n]$ is stationary, so that $e[n]$ and $r[n]$, which are derived as time-invariant (but nonlinear) functions of $x[n]$, are likewise stationary [14, p. 238]. Let the power spectra of $e[n]$ and $r[n]$ be denoted $S_{ee}(\omega)$ and $S_{rr}(\omega)$, respectively, where ω is radian frequency. In terms of these power spectra, (2) can be rewritten [14, p. 292]

$$S_{rr}(\omega) \approx |G(e^{j\omega})|^2 S_{ee}(\omega); \quad -\pi \le \omega < \pi. \tag{3}$$

Now it is assumed that both prefilter and postfilter are approximately all-pass, meaning that their magnitude-frequency responses are nearly flat over the full spectrum. Consistent with this, it can be assumed without loss of generality that $|G(e^{j\omega})| = 1$ for $-\pi \le \omega < \pi$ (whatever scaling factor is needed to make this true can be canceled by an appropriate gain in the prefilter). Thus, (3) reduces to

$$S_{rr}(\omega) \approx S_{ee}(\omega), \tag{4}$$

Which implies that the mean-square value of $r[n]$ is nearly equal to that of $e[n]$. That is, the system's mean-square error is nearly equal to that incurred by quantizing the intermediate (prefiltered) signal $u[n]$.

4. Statistical characterization of the intermediate signal

Since each sample in $u[n]$ is the (weighted) sum of *independent* random variables, its probability density function is approximately Gaussian, provided that the sum includes a sufficient number of variables, each with nonnegligible but not disproportionately large weight (by Liapounov's central limit theorem [15, p.200]). These conditions will be satisfied when the impulse response of the prefilter, $h[n]$, has significantly nonzero values distributed over a sufficiently long interval. In this paper, a filter with this property will be called time-dispersive.

The most convenient measure of the extent to which the PDF of the source is modified is the observed performance of simple quantization of the intermediate signal under the assumption of Gaussian distribution. It has been found experimentally that for a Laplacian source, FIR time-dispersive prefilters and postfilters of length 30 are sufficiently long to yield a signal-to-quantization-error ratio within 0.2 dB of the best possible (that for a Gaussian source) at rates up to 5 bits per sample. In the case of a gamma source, a length of 120 is required for the same level of performance. However, even when much shorter filters are used, a significant improvement over direct quantization results (see Section 6). Design of appropriate prefilters and postfilters of a given length is discussed in Section 5.

Another way to gauge the extent to which the PDF of the input is modified is to examine histograms. Figure 3 shows histograms based on 10,000 samples from simulated sources, before and after prefiltering.

A legitimate objection to the foregoing analysis is that in many applications, the assumption of independence of successive source samples is unjustified, so that prefiltering may not make the PDF approximately Gaussian. In fact, it is easy to construct a source for which prefiltering makes the distribution appear *less* Gaussian — for example, simply filter a gamma source by $G(z)$, and use the result as $x[n]$. In all such cases, the previous analysis can be made to apply if the source samples are rearranged or scrambled in a pseudorandom manner prior to prefiltering, and subsequently restored to their original ordering after postfiltering. Note that scrambling and inverse scrambling are linear and energy-preserving operations, so that the analysis of Section 3 applies, and quantization error is preserved. However, scrambling and inverse scrambling necessarily introduce considerable delay, and therefore may not be appropriate in some applications.

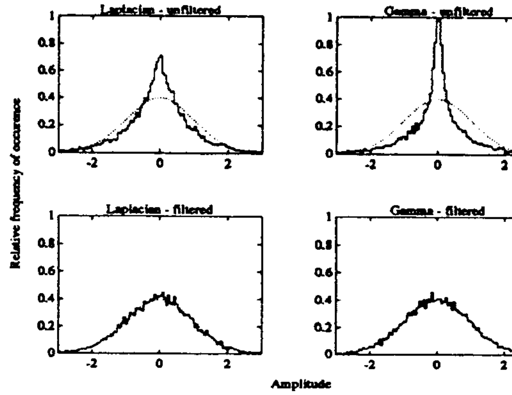

Figure 3: Histograms of original and filtered sources, based on 10,000 pseudorandom samples. A 31-tap filter was used in the Laplacian case, while a 60-tap filter was used in the gamma case. For reference, an appropriately scaled version of the Gaussian PDF is also shown (dotted curve).

5. Design of time-dispersive prefilters and postfilters

It is desired that the prefilter and postfilter be approximate inverses of one another, that their impulse responses have envelopes that extend sufficiently over time, and that their magnitude-frequency responses be approximately flat. One approach to obtaining such filters is to begin with "initial guess" filters that have *roughly* the right properties, then refine these by numerical optimization. In particular, the following procedure has proven to be successful.

Begin with a windowed "chirp" signal (swept sinusoid) as an initial guess for the impulse response of the prefilter, and use the same chirp, but time-reversed, as the initial guess for the impulse response of the post-filter. To simplify notation in the present section, the postfilter is not required to be causal, and the delay D of the cascade is taken to be zero. Each of the two initial guess filters has the desired property that the energy in its impulse response is distributed over the entire region of support; that is, the filters are time-dispersive. In order to ensure that the remaining requirements are met — that the prefilter and postfilter be approximate inverses of each other, and that each have an approximately flat magnitude-frequency response — a numerical procedure is used to modify the initial guess filters to minimize the total square difference between the convolution of $h[n]*g[n]$ and the unit-sample sequence $\delta[n]$. That is, a local minimum is sought of the objective function

$$E = \sum_{n=-\infty}^{\infty} \left[\sum_{k=-\infty}^{\infty} h[k]g[n-k] - \delta[n] \right]^2 \tag{5}$$

over the joint space of prefilter and postfilter coefficients $\{h[n], g[n]\}$, beginning the search at the specified initial guess.

Observe that the initial guess filters, being time-reversed versions of each another, have identical magnitude-frequency responses. By maintaining this relationship throughout the optimization, so that the optimized filters also end up as time-reversed versions of each other, the magnitude-frequency response of each optimized filter can be made to be approximately flat. This follows because the magnitude-frequency response of the cascade — which is the product of the individual responses — must be flat if the two filters are to be inverses of each another. The time-reversed relationship can either be maintained explicitly by adding a simple constraint (i.e., optimizing over only one of the filters and fixing the other according

1990 International Symposium on Information Theory and Its Applications (ISITA'90)
Hawaii, U.S.A., November 27-30, 1990

to the time-reversed relationship), or else the symmetry of the objective function can be relied upon to maintain the relationship from the initial guess. The latter approach was found to work consistently in the present investigation.

It is natural to question the existence of local minima, convergence issues, and so on; however, such a formal treatment of the optimization problem is avoided here, on the grounds that in practice, a local minimum seems to be obtainable quickly and consistently using any of a variety of well-known optimization procedures.

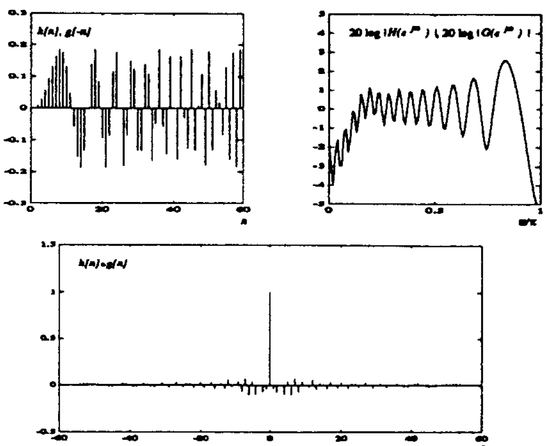

Figure 4: Prefilter and postfilter characteristics *before* optimization (see Section 5).

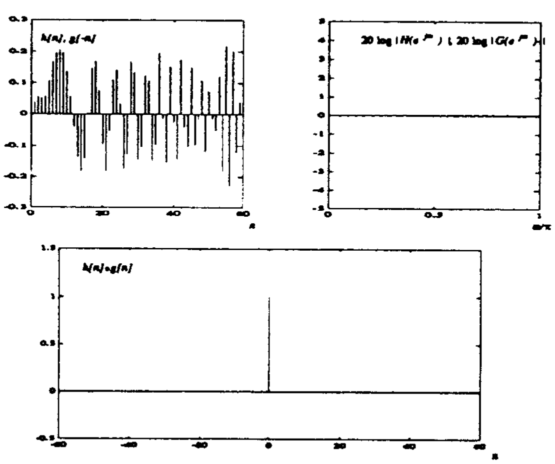

Figure 5: Prefilter and postfilter characteristics *after* optimization (see Section 5).

Figures 4 and 5 show the characteristics of the prefilter and postfilter in a 60-tap design example, before and after optimization, respectively. Also shown in each figure is the convolution of $h[n]$ and $g[n]$ and the corresponding magnitude-frequency response of the cascade of the two filters. In this example, the chirp initial guess filters were taken to be

$$h[n] = g[-n] = A\sin(\pi(n^2 - n)/120), \quad n = 0, \ldots, 59,$$

where A was chosen to make the sum of the squares of the coefficients in each filter unity. Note that the optimized filters (Figure 5) have the desired properties: the prefilter and postfilter are approximate inverses of each other, have nearly flat magnitude-frequency responses, and have impulse responses with significant energy distributed well over the entire region of support.

6. Experimental results

The dependence of the performance of the proposed system on the length of the filters is illustrated in Figure 6, for the range of 5-37 taps. To obtain each point in the graphs. the sources were simulated using techniques described by Knuth [16, vol 2, pp. 128-129], and the performance of the proposed quantization system was measured for 10,000 samples. Observe that for both sources, even very short filters yield a considerable improvement in performance. Although not shown in the figure, it was found that as the filters are made longer than 37 taps, the improvement in performance continues to be noticeable for the gamma source, but not for the Laplacian source.

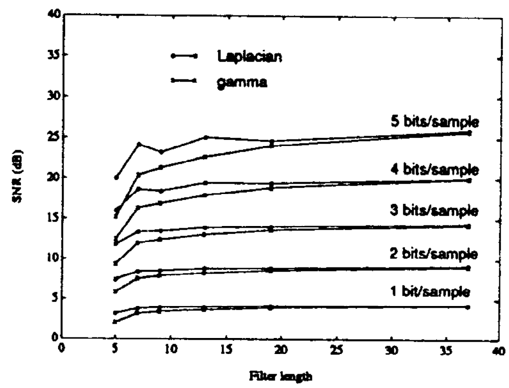

Figure 6: Experimentally determined dependence of performance of the proposed system on the length of the filters. The procedure described in Section 5 was used to design the filters.

In obtaining the remaining experimental results presented in this section, filters of length 31 and 60 were used in the Laplacian and gamma cases, respectively. Also, unless otherwise stated, all measurements were based on 10,000 samples.

Figure 7 shows mean-square error as a function of the number of quantization bits for simulated Laplacian and Gamma sources, for both simple quantization and the filter-based quantization scheme. Also shown are samples of the rate-distortion functions for these two sources. Note that the performance is, as expected, approximately that of simple quantization of a Gaussian source. The improvement over quantization without filtering is particularly significant at low bit-rates; in fact, by comparing the results with those presented in [4], it can be concluded that at 1 bit/sample, the improvement over direct quantization obtained by prefiltering and postfiltering is roughly the same as would be obtained by three-dimensional vector quantization.

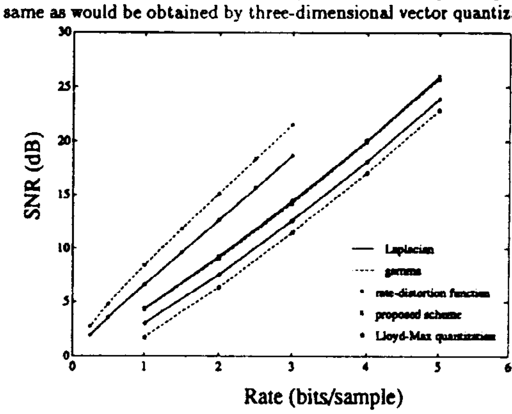

Figure 7: Experimental performance (in terms of mean-square error) of the proposed scheme relative to direct Lloyd-Max quantization for a Laplacian and a gamma source. Also shown on each graph are three samples of the corresponding rate-distortion function, computed via the Blahut algorithm [10].

Besides reduced mean-square error, an important advantage of the proposed technique over simple quantization is its robustness to

1990 International Symposium on Information Theory and Its Applications (ISITA'90)
Hawaii, U.S.A., November 27-30, 1990

errors in modeling the input. Specifically, the prefiltering operation will tend to make *any* memoryless source appear more Gaussian, so that the performance of the system does not depend critically on accurate modeling of the input PDF. Figure 8 illustrates the relative insensitivity of the proposed technique to modeling errors. Shown is the performance of the proposed system and that of direct quantization, when each source is mistakenly modeled as the other. Note that the performance of direct quantization is reduced because of the mismatch, while that of the proposed system is unaffected. This implies that the proposed system can be used with some confidence even when relatively little is known about PDF of the source. In some applications, this robustness is more significant than reduction of mean-square error. The most notable prior work in robust simple quantization is by Bath and Vandelinde [17]. In their approach, a minimum level of MSE performance is guaranteed so long as the input PDF belongs to a certain class; however, that performance is considerably worse than the performance of Lloyd-Max quantization of a Gaussian source. In contrast, for the same class of input PDF, the performance of the quantization scheme proposed in this paper is always *roughly equal* to that of Lloyd-Max quantization of a Gaussian source.

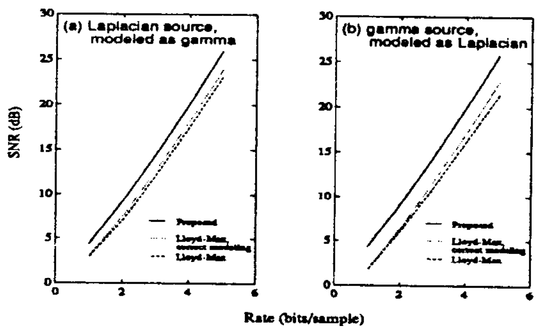

Figure 8: Experimental mean-square error performance of the proposed system and that of direct quantization in the case of quantizer mismatch. For reference, also shown is the mean-square error of direct quantization when the input is correctly modeled (dotted line). In (a) the source is Laplacian distributed, but modeled as gamma. In (b) the source is gamma distributed, but modeled as Laplacian.

Another potential advantage of the proposed system is reduction of correlation between input and quantization error at low bit rates. Preliminary experiments have revealed that the reduction can be significant at low bit rates, where correlation is usually most objectionable. However, further experimentation is required to determine the precise extent of the reduction, as well as its benefit in practical coding systems.

7. Conclusions

A technique to improve the performance of simple quantization for memoryless sources has been proposed, and experimental results have been presented that show that in practice the system performs as expected. The technique results in a significant reduction in mean-square quantization error and offers relative insensitivity to errors in modeling the input distribution.

References

[1] Paez, M.D., and Glisson, T.H., "Minimum mean squared error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.*, vol. COM-20, no. 4, April 1972, pp. 225-230.

[2] Farvardin, N., and Modestino, J., "Optimium quantizer performance for a class of non-Gaussian memoryless sources," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 3, May 1984, pp. 485-497.

[3] Fehn, H. G., and Noll, P., "Multipath search coding of stationary signals with applications to speech," *IEEE Trans. Commun.*, vol. COM-30, no. 4, April 1982, pp. 687-701.

[4] Fischer, T. R., and Dicharry, R. M., "Vector quantizer design for memoryless Gaussian, gamma, and Laplcian sources," *IEEE Trans. Commun.*, vol. COM-32, no. 9, Sept. 1984, pp. 1065-1069.

[5] Jayant, N. S., and Noll, P., *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

[6] Rabiner, L.R., and Schafer, R., *Digital Coding of Speech Signals*, Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[7] Gallager, R. G., *Information Theory and Reliable Communication*, New York: Wiley, 1968.

[8] Lloyd, S. P., "Least squared quantization in PCM," unpublished memorandum, Bell Lab., 1957; reprinted in *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, March 1982, pp. 129-137.

[9] Max, J., "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-6, no. 2, March 1960, pp. 7-12.

[10] Blahut, R.E., "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 4, July 1972, pp. 460-473.

[11] Viterbi, A. J., and Omura, J.K., "Trellis encoding of memoryless discrete-time sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 3, May 1974, pp. 325-332.

[12] Strube, H. W., "How to make an all-pass filter with a desired impulse response," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 2, April 1982, pp. 336-337

[13] Zenith Electronics Corporation, "Zenith spectrum compatible HDTV system," Glenview, Illinois, 1988.

[14] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, second ed., New York: McGraw-Hill, 1984.

[15] Thomas, J. B., *An Introduction to Applied Probability and Random Processes*, New York: Wiley, 1971.

[16] Knuth, D.E., *The Art of Computer Programming*, 2nd ed., vol. 2, Reading, Massachusetts: Addison-Wesley, 1981.

[17] Bath, W.G., and Vandelinde, V.D., "Robust memoryless quantization for minimum signal distortion," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, March 1982, pp. 296-297.

Ashok Chhabedia Popat was born in Lynn, Massachusetts on January 28, 1961. He received the S.B. and the S.M. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, in 1986 and 1990, respectively. While an undergraduate, he worked part-time at Codex Corporation in Canton, Massachusetts on signal processing problems relating to voiceband modem operation and speech- and image-coding. Between 1986 and 1988, he worked at Hewlett-Packard in Rohnert Park, California, as an applications engineer. Between 1988 and 1990, he worked on coding for high-definition television as a research assistant at the MIT Advanced Television Research Program. In 1990 he began his doctoral studies at the Ecole Polytechnique Fédérale de Lausanne, Switzerland. His research interests include digital signal processing, information theory, optimization, and statistical inference.

Kenneth Zeger was born in Boston, Massachusetts on August 18, 1963. He received both the S.B. and S.M. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT) in 1984, and the M.A. degree in pure mathematics and the Ph.D. in electrical and computer engineering at the University of California, Santa Barbara (UCSB) in 1989 and 1990 respectively. During 1980-81 he worked on adaptive antenna array designs for Z-A Inc. in Glenside, PA. He has worked on real-time speech recognition for Hewlett-Packard Co. at the General Systems Division in Sunnyvale, CA and on speech compression techniques at HP Laboratories in Palo Alto, CA during the years 1982-1985. In 1984 he served as a consultant to Automatic Data Processing Co. on digital network design. In July 1990 he joined the Electrical Engineering faculty at the University of Hawaii as an Assistant Professor. His present research interests include combined source/channel coding, speech and image compression, and computational complexity theory. Dr. Zeger was awarded a four-year Faculty Development Graduate Fellowship by the American Electronics Association in 1985 and was the recipient of a University of California Regents Fellowship in 1989. He is a member of the Communication Theory Technical Committee of the IEEE Communications Society, and served as Co-Chairman of the 1990 IEEE Workshop on Communication Theory, held in Ojai, California.