ON OPTIMAL AND MINIMUM-ENTROPY DECODING

W. Bastiaan Kleijn[*]

Department of Electrical Engineering
Delft University of Technology
Delft, The Netherlands

ABSTRACT

In the quantization of a signal in speech coding, dependencies between its samples are often neglected. Generally, these dependencies are then also neglected at the decoder. However, usually *a priori* information about these dependencies is available, making it possible to improve decoder performance by means of enhanced decoding. An attractive feature of enhanced decoding is that it can be applied to existing coding standards. This paper describes several enhanced decoding methods, including a vector decoding method and a method which aims at reducing the differential entropy rate of the decoded signal. Experimental results are used to confirm that both these decoding procedures can provide better performance than conventional decoding for common signal/encoder combinations.

1 INTRODUCTION

The quantizers in speech coders operate on scalar or vector signals. We define as a *sample* the finite sequence of vectors or scalars which is quantized in one quantization operation (this sequence may contain only one scalar or vector). Quantizers often operate on each sample independently of the other samples. Inherently, coding efficiency is lost in such an encoding process. This loss in coding efficiency is particularly strong when the unquantized samples are highly dependent, *i.e.*, when the signal has a high level of redundancy. The losses in efficiency are often the result of practical constraints on computational complexity, delay, robustness to channel errors, and compatibility with older standards.

It is natural to create a decoder which reflects the design of the encoder. That is, the decoder considers only one sample at a time. However, as will be shown in this paper, particularly when there is a high level of redundancy in the encoded signal which is not exploited during encoding, it may be possible to exploit such dependencies at the *decoder*. Sufficient *a priori* information concerning the dependencies is often known to enhance the output significantly. Furthermore, the forementioned complexity, robustness, and compatibility constraints often do not apply to the decoder. Thus, enhanced decoding procedures which exploit sample dependencies are of practical significance.

An obvious application of enhanced decoding is the combination of a relatively coarse scalar quantizer and a signal which is generally smooth, but which may include regions of rapid change. This situation is typical for coded speech

[*]current address:
Department of Speech, Music, and Hearing
KTH (Royal Institute of Technology)
Box 700 14, 100 44 Stockholm, Sweden

parameters such as the line spectral frequencies, the speech signal power, and the pitch period.

This paper has the following outline. Section 2 provides a description of three decoding aproaches: vector decoding, set-theoretic estimation, and constrained optimization. It also contains a few experimental results for the vector decoding procedure. Section 3 then goes into more detail on a particular constrained optimization procedure, which is referred to as minimum-entropy decoding. Conclusions are provided in section 4.

2 METHODS FOR DECODING

A quantization index i received at the decoder indicates in which quantizer cell (Voronoi region) the original sample value is located. The requirement that the decoded value should fall within this Voronoi region will be called the quantization consistency constraint. In most decoders this constraint is satisfied. Usually the decoded value is the optimal decoded value (or an approximation thereof), given the quantization index i, in the least squares sense. If $Q_{\text{opt}}^{-1}(\cdot)$ denotes this optimal decoder, then [1]

$$Q_{\text{opt}}^{-1}(i) = E[s|i], \qquad (1)$$

where $E[s|i]$ is the expectation value of the original sample s given that the quantization index is i. Obviously, this single-sample decoding method does not account for *a priori* information about sample dependencies which might be available to the decoder. In the following subsections three different approaches to the consideration of sample dependencies are described.

2.1 Vector decoding

The vector decoding method is a straightforward generalization of the above procedure, where the decoder now uses simultaneously multiple quantization indices. Despite its conceptual simplicity, the method does not appear to be in use. Let $s = [s(0), s(1), \cdots, s(N-1)]$ denote a sequence of N samples and let $\mathbf{i} = [i(-q), i(-q+1), \cdots, i(N-1+q)]$ $(q \geq 0)$ denote a sequence of $N + 2q$ quantization indices. The integer q represents the length of the "overhang" of the index sequence beyond the interval of the sample sequence. The optimal decoder for this generalized case is

$$Q_{\text{opt}}^{-1}(\mathbf{i}) = E[\mathbf{s}|\mathbf{i}]. \qquad (2)$$

An asymptotically optimal decoder $Q^{-1}(\cdot)$ can be obtained by means of averaging over an appropriately collected, large data base with data vectors $\mathbf{s}_k$ (the index k identifies the vector within the data base). Let $A(\mathbf{i})$ denote the set of data vectors $\mathbf{s}_k$ with quantization vector $\mathbf{i}$:

$A(\mathbf{i}) = \{\mathbf{s}_k : Q(\mathbf{s}_k) = \mathbf{i}\}$ and let $||A(\mathbf{i})||$ denote the number of entries in the set $A(\mathbf{i})$. Then, a good decoder is

$$Q^{-1}(\mathbf{i}) = \frac{1}{||A(\mathbf{i})||} \sum_{\mathbf{s}_k \in A(\mathbf{i})} \mathbf{s}_k. \qquad (3)$$

An apparent weakness of vector decoding is the remaining division of the sampled signal into consecutive sequences and the resulting loss in performance at the boundaries of these sequences. However, this is only the case for $q = 0$. By choosing q sufficiently large (*i.e.*, by including indices corresponding to samples beyond the sequence itself), boundary effects are removed. Note that the case where the decoded sequence contains only one sample ($N = 1$) can also provide enhanced decoding for $q > 0$.
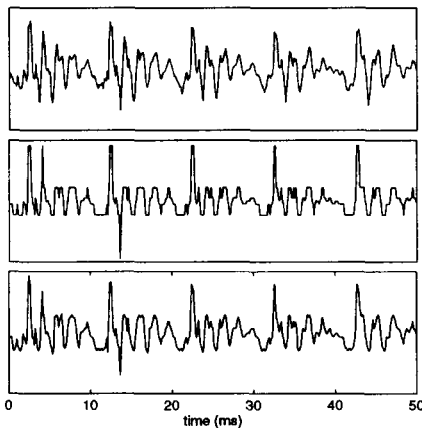

Figure 1. Top to bottom: the original signal,
output of scalar 3-bit μ-law encoder with scalar decoder,
output of the same encoder and 4-sample vector decoder.

A more serious drawback of vector decoding is the storage requirement for the codebook. The codebook size increases exponentially with the number of samples in the index sequence, restricting the length of the sequences. For a bit rate of r bits per sample, the codebook size is $2^{r(N+2q)}$. It may be possible to reduce the codebook size by only storing common vectors and use scalar decoding for rare vectors.

Vector decoding should be useful in practical applications with low bit allocations per sample. Obviously, the method works particularly well for signals which are of restricted structure (*e.g.*, DTMF tones). However, the method also provides usefully improved performance for practical signal-encoder combinations, such as scalarly quantized speech-model parameters and scalar quantized speech. Figure 1 shows an example of a scalar 3 bit/sample encoding with a conventional scalar encoder (based on down-sampling the G.711 μ-law codebook) coder and a four-dimensional vector decoder ($q = 0$). Table 1 shows the performance in terms of segmental SNR (again, $q = 0$). The testing and training data bases were each drawn from different data bases (obtained from different sources), and contained eight speakers (total 1 minute) and twelve speakers (total 2 minutes), respectively. Signal segments with an RMS of less than 100 in the 16-bit resolution signal were omitted from the segmental SNR measurements.

2.2 Set-theoretic estimation

A result of the usage of optimal encoders is that the Voronoi regions form convex sets [1]. In fact, each Voronoi region

Table 1. Vector decoder performance for μ-law coded speech

bit/sample	3	3	3	4	4
decoder dimension	1	3	4	1	3
SSNR (dB)	7.81	8.13	8.52	13.08	13.80

can itself be seen as an intersection of convex sets, which are determined by hyperplanes. Feasibility algorithms are algorithms which, given a number of (closed) convex sets which have a nonempty intersection, find a point which lies within the intersection of the convex sets [2]. Thus, these algorithms find a nonunique solution which is consistent with a set of constraints which are formulated as convex sets. Using only the same information as a conventional decoder, the feasibility algorithms would simply find some point within the Voronoi region.

The strength of set-theoretic estimation is that additional convex sets can be specified. These sets must be consistent with the existing sets and with each other, *i.e.*, the intersection of all convex sets specifying the problem should not be empty. The additional convex sets usually represent explicit *a priori* information such as a description of the smoothness of the original signal (*e.g.*, [3]).

A good example of the power of set-theoretic estimation is an oversampling D/A converter. In this case, the constraints are that the signal samples are within the appropriate Voronoi regions and that the signal is bandlimited [4]. The straightforward approach of conventional quantization and decoding followed by a low-pass filtering operation generally violates the first condition and performance can be improved significantly using set-theoretic estimation.

For a practical application in speech coding, set-theoretic estimation of the decoded value does have drawbacks. The computational complexity of the feasibility algorithms, which are iterative algorithms based on projections, is generally high. The additional information must be formulated in terms of a consistent convex set, which sometimes requires additional information to be transmitted. Finally, in the case that a proper criterion for optimization of the decoded value within a Voronoi region can be defined (which is not formulated for computational convenience), then the set-theoretic approach is at a disadvantage.

2.3 Criterion-based decoding

It is sometimes natural to introduce *a priori* information by means of a criterion which is easily optimized under the constraint that the solution remains within the appropriate Voronoi region. In contrast to the set-theoretic approach, this will generally result in a unique solution. An example of the criterion-based approach is the decoding method for line spectral frequencies described in [5]. There, a newly-defined smoothness criterion was maximized under the constraint that each sample falls within the Voronoi regions specified by the quantization index.

Other decoding criteria can be defined. In the next section, the operation of a new criterion will be described. It is based on the fact that the mutual information between the signal samples decreases when the samples are quantized independently. This can be interpreted as a decrease in the structure of the signal during the quantization process. Thus, a natural criterion for enhanced decoding is to minimize the differential entropy rate subject to the quantization consistency constraint. This general strategy is useful for signal which have a relatively high level of redundancy, such as speech and speech-model parameters in combination with a quantizer which does not exploit this redundancy.

3 MINIMUM-ENTROPY DECODING

The minimum-entropy decoding procedure is aimed at improving the decoder performance when the original samples have a strong interdependency which is not exploited during the encoding process. The method is motivated by a property of the quantization process. When a signal sample is quantized, it is assigned to a particular Voronoi region. As will be shown in subsection 3.1, the quantization process increases the uncertainty about the signal, *i.e.*, it increases the differential entropy rate of the signal. The present approach towards improved decoding aims to undo this increase in the differential entropy rate.

As an extreme example, consider the quantization of a deterministic signal, such as a sinusoid. (Voiced speech signals are sometimes described as being similar to deterministic signals [6].) The differential entropy rate of these signals is negative infinite and upon quantization the differential entropy rate becomes a finite value which is a function of the size of the Voronoi regions and the probability of their selection. In other words, the optimal decoded signal has the minimum differential entropy rate consistent with the quantization consistency constraint. This leads to the notion that for such signals minimization of the differential entropy rate under the quantization consistency constraint might be a good decoding strategy. However, additional constraints are necessary since minimum differential entropy solutions are, in general, not unique. (For example, for a one-dimensional sample a mapping which reflects each Voronoi region around its center does not affect the differential entropy). This can be resolved by selecting a particular method to reduce the differential entropy. The method described in subsection 3.2 is a reasonable choice since it reduces the differential entropy *efficiently*, *i.e.*, the modification in the signal associated with the differential entropy reduction is minimal.

Of course, the above motivation for decoding by means of constrained minimization of the differential entropy is less compelling for signals consisting of highly independent samples. Thus, the performance of the algorithm can be expected to vary with the input signal. However, since most coded signals exhibit structure, improved performance can be expected from an efficient reduction the differential entropy rate for many signal/encode combinations.

3.1 Differential entropy and quantization

Consider the conditional differential entropy of a sample x, given a vector of samples $\mathbf{y}$, $h(x|\mathbf{y})$. Of particular interest is the case where the elements of $\mathbf{y}$ form an ordered sequence of samples of a stationary signal, and x is the next sample of this sequence. In the limit where the dimensionality of $\mathbf{y}$ approaches infinity, the conditional entropy approaches the differential entropy-rate of the signal. In this subsection, the effect of quantization on $h(x|\mathbf{y})$ will be discussed.

Prior to considering the effect of quantization on $h(x|\mathbf{y})$, it is useful to define a meaning for differential entropy of the quantized signal. Here, the differential entropy $h(k)$ of the quantized signal x is defined as the differential entropy of a signal which has uniform probability density within a Voronoi region k, with the total probability within each Voronoi region, denoted by $p(k)$, being identical to that prior to quantization. The relation between the entropy of the quantization index k, denoted $H(k)$, and the differential entropy $h(k)$ is easily derived and is found to be

$$h(k) = H(k) + \sum_k p(k) \log[\mathbf{A}_k], \quad (4)$$

where $\mathbf{A}_k$ is the volume of the Voronoi region.

The effect of quantization on the conditional entropy $h(x|\mathbf{y})$ can now be considered in more detail. To distinguish between effects which are single-sample in nature and effects which are related to the dependencies between x and $\mathbf{y}$, the differential conditional entropy is written in terms of a first-order differential entropy and a mutual information:

$$h(x|\mathbf{y}) = h(x) - I(x;\mathbf{y}). \quad (5)$$

It is natural that the conditional entropy of x given $\mathbf{y}$ increases when quantization is applied to the samples of $\mathbf{y}$, independent of x. One expects quantized data to carry less information about future data than unquantized data. This result is straightforward to prove. Let $\mathbf{i}$ denote the vector of indices resulting from independent quantization of the sample elements of $\mathbf{y}$. Each $\mathbf{i}$ specifies a Voronoi region in the $\mathbf{y}$ space with volume $\mathbf{A_i}$. Using Jensen's inequality [7] and the fact that a function $t \log t$ of t is convex, it is simple to show that quantization of $\mathbf{y}$ decreases the mutual information between x and $\mathbf{y}$:

$$
\begin{aligned}
I(x;\mathbf{y}) &= \int_x \int_{\mathbf{y}} f(x,\mathbf{y}) \log\left[\frac{f(x,\mathbf{y})}{f(x)f(\mathbf{y})}\right] d\mathbf{y}\,dx \\
&= \int_x \sum_{\mathbf{i}} p(\mathbf{i}) \int_{\mathbf{A_i}} f(x|\mathbf{y})\frac{f(\mathbf{y})}{p(\mathbf{i})} \log\left[\frac{f(x|\mathbf{y})}{f(x)}\right] d\mathbf{y}\,dx \\
&\geq \int_x \sum_{\mathbf{i}} \int_{\mathbf{A_i}} f(x,\mathbf{y})d\mathbf{y} \log\left[\frac{\int_{\mathbf{A_i}} f(x,\mathbf{y})d\mathbf{y}}{f(x)p(\mathbf{i})}\right] dx \\
&= \int_x \sum_{\mathbf{i}} f(x,\mathbf{i}) \log\left[\frac{f(x,\mathbf{i})}{f(x)p(\mathbf{i})}\right] dx \\
&= I(x;\mathbf{i}). \quad (6)
\end{aligned}
$$

The increase in the conditional differential entropy of x given $\mathbf{i}$ due to quantization of x also affects the mutual information between x and $\mathbf{i}$ (or $\mathbf{y}$). Let k denote the quantization index of x, then

$$I(x;\mathbf{i}) \geq I(k;\mathbf{i}). \quad (7)$$

where the same principles as in equation 6 were used.

Because of the assumption of uniform probability density within the Voronoi region, the differential entropy of the quantized parameter x is greater or equal than the differential entropy of the unquantized parameter x (this can again be shown with Jensen's inequality):

$$h(k) \geq h(x). \quad (8)$$

Note that inequality 8 describes the effect of quantization on the differential entropy of a sample without accounting for dependencies between the samples.

In summary, it has been shown that the mutual information between x and $\mathbf{y}$ decreases both when the samples of $\mathbf{y}$ are quantized independently and when x is quantized independently. This can be interpreted as a decrease in the structure of the signal when it is subjected to independent quantization of its samples. In addition, using the assumption of uniform density within the Voronoi region, the first-order differential entropy of x increases when it is quantized independently. Both of these contributions lead to an increase of the differential entropy rate of the signal when the signal samples are quantized independently.

3.2 Differential entropy reduction

In this subsection a practical method to perform differential entropy minimization of the signal is described. Consider a vector s describing a finite signal segment. Let us gather all but one of the elements of s (in arbitrary order) to form a vector $\mathbf{v}$, and denote the remaining element by w. A predictor $g(\mathbf{v})$ predicts the sample value w with error e:

$$e = w - g(\mathbf{v}). \tag{9}$$

The goal is to minimize the differential entropy rate of the signal. We move towards this goal with minimizing the differential entropy of the vector s, $h(\mathbf{s}) = h(w, \mathbf{v})$, by modifying w. The procedure can then be iterated over the samples of s. It is convenient to express the differential entropy as

$$h(w, \mathbf{v}) = h(\mathbf{v}) + h(w) - I(w; \mathbf{v}). \tag{10}$$

Using equation 9 it can be shown that

$$I(w; \mathbf{v}) = h(w) - h(e) + I(e; \mathbf{v}). \tag{11}$$

Combining equations 10 and 11 gives

$$h(w, \mathbf{v}) = h(\mathbf{v}) + h(e) - I(e; \mathbf{v}). \tag{12}$$

The differential entropy $h(w, \mathbf{v})$ can be reduced by scaling the prediction error e by a factor α, with $|\alpha| < 1$. While $I(e; \mathbf{v})$ is insensitive to the scale of e,

$$h(\alpha e) = h(e) + \log(|\alpha|). \tag{13}$$

Note that the scaling of the parameter e changes the parameter w into $w - (1 - \alpha)e$. Thus, it is seen that choosing the optimal predictor in the entropy-reduction algorithm results in the smallest modification of the signal when decreasing the entropy. Furthermore, it is noted that the addition of a quantization consistency constraint does not affect the basic functioning of the algorithm.

3.3 Results

To illustrate the performance of the minimum-entropy method, it was applied to μ-law coded speech and tone signals, sampled at 8 kHz. A linear interpolator with 8 samples on each side of the interpolated sample was used for the predictor $g(\cdot)$. The interpolator was optimized for the conventionally decoded data. (This is not strictly correct. The predictor should be optimized for the samples being uniformly distributed within the Voronoi regions. For the linear predictor, the present procedure was adequate.) The interpolator was updated every 10 samples and it was not changed during the iterative process described below. To lower computational complexity, the interpolated values are computed simultaneously for the 10-sample speech segments. Let j denote the sample time index, $i(j)$ the sample quantization index and $s_m(j)$ the decoded sample value for iteration m. Furthermore, let $l(i(j))$ denote the lower limit of the Voronoi region with index $i(j)$, $u(i(j))$ the corresponding upper limit, and $\hat{s}_m(j)$ the output of the interpolator. Starting from the conventionally decoded values, the entropy is reduced by iteratively scaling the prediction error $s_m(j) - \hat{s}_m(j)$ by a factor α slightly less than unity:

$$s_{m+1}(j) = \\ \min(\max(s_m(j) - (1 - \alpha)(s_m(j) - \hat{s}_m(j)), l(i(j))), u(i(j))) \tag{14}$$

Experimental results for μ-law coded speech are shown in Table 2. The results were obtained with five iterations. Usually the performance degraded under further minimization of the differential entropy rate, which is consistent with the original data not being deterministic. The measurement and data base are identical to that used for testing vector decoding (Table 1). The improvements are larger for voiced than unvoiced speech segments, which is consistent with the former being more similar to a deterministic signal.

The results for sinusoids are not reported. As expected, the decoding method performs very well for such deterministic signals, and enhancement by up to 10 dB was obtained.

Table 2. Segmental SNR (SSNR) for conventional and minimum entropy (ME) decoding of μ-law coded speech

bit/sample	3	3	4	4	5	5
decoder	conv.	ME	conv.	ME	conv.	ME
SSNR (dB)	7.81	9.44	13.08	15.10	18.92	20.48

4 DISCUSSION

It has been shown that if a quantizer does not exploit the dependencies between samples of a signal, then these dependencies often can be exploited using *a priori* information in the decoder. Three basic classes of enhanced decoding were described: vector decoding, set theory based estimation, and methods using an optimization criterion. Much literature on the set theory based methods exists.

The vector decoding method and a method which reduces the differential entropy efficiently under the constraint that each decoded value remains in the proper Voronoi region were subjected to experiments. The results suggest that vector decoding can lead to a significant increase in performance over conventional decoders for common quantizer/signal combinations. The disadvantage of this method is the storage requirement. While the entropy minimization method is not asymptotically optimal like the vector decoding method, it was was shown to perform better than the vector decoder in the practical experiments performed.

In contrast to conventional signal enhancement techniques, enhanced decoding exploits knowledge about the quantizer cells. Thus, good enhanced decoding techniques should perform better than a conventional decoder followed by a separate enhancement technique. However, no comparisons with existing conventional signal enhancement techniques were described in the present paper.

REFERENCES

[1] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Dordrecht, Holland: Kluwer Academic Publishers, 1991.

[2] P. L. Combettes, "The foundations of set theoretic estimation," *Proc. IEEE*, vol. 81, pp. 182–208, 1993.

[3] K. Asai, N. T. Thao, and M. Vetterli, "A study of convex coders with an application to image coding," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, (Adelaide), pp. V581–V584, 1994.

[4] N. T. Thao and M. Vetterli, "Consistent signal reconstruction and convex coding," in *SVD and signal processing, III* (M. Moonen and B. De Moor, eds.), pp. 79–95, Elsevier Science Publishers, 1995.

[5] H. P. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Detroit), pp. 732–735, 1995.

[6] G. Kubin, "Nonlinear processing of speech," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 557–610, Elsevier Science Publishers, 1995.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.