1994 International Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong

Vector Quantization View of $\Sigma\Delta$ Modulation

Nguyen T. Thao

Department of EEE
Hong Kong University of Science & Technology

Martin Vetterli

Department of EECS
University of California, Berkeley

Abstract

We analyze the behavior of a single-loop $\Sigma\Delta$ modulator using a vector quantization (VQ) approach. We extract the encoder-decoder structure typical of VQ, existing in a $\Sigma\Delta$ modulator. The study of the encoding part gives the intrinsic behavior of the modulator. Bounds to the intrinsic performance of the $\Sigma\Delta$ modulator can be derived through this approach in the oversampling situation, assuming periodicity of the bandlimited input signals.

1 Introduction

Substantial research has been devoted to the theoretical analysis of $\Sigma\Delta$ modulation, involving statistical approaches [1], as well as nonlinear dynamics approaches [2]. In this paper, we propose to study the behavior of a single-loop $\Sigma\Delta$ modulator through a vector quantization (VQ) point of view. Indeed, the input of a $\Sigma\Delta$ modulator is an N-point sequence of samples which can be considered as a vector of $\mathbf{R}^N$. We show that the output of a modulator can be interpreted as the codevector output of a vector quantizer. As in the general case of VQ, we show that a $\Sigma\Delta$ modulator has a built-in encoder-decoder structure. Because the decoding part can always be modified by postprocessing, we explain that the intrinsic behavior of the modulator lies in its encoding part and consequently in the vector space partition induced by the encoding part. Assuming no overloading of the scalar quantizer, we study this partition both when input sequences are not constrained and in the oversampling situation. In the second case, we confine ourselves to periodic bandlimited input signals. We detail the particular structure of the partition and derive that the mean squared error (MSE) of the $\Sigma\Delta$ modulator with optimal decoding is asymptotically lower bounded with respect to the oversampling ratio R in $\mathcal{O}(R^{-4})$.

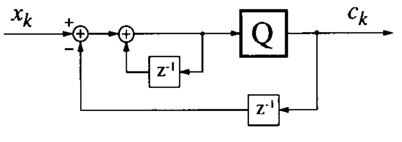

Figure 1: Block diagram of the single-loop $\Sigma\Delta$ modulator.

2 Vector quantization view

Figure 1 shows the block diagram of a single-loop $\Sigma\Delta$ modulator, where Q symbolizes a scalar quantizer. In general, the input is discrete-time and of finite length as it has to be the case in reality. Let N be the total number of input samples and $x_1, ..., x_N$ be the samples. Then, the output is an N-point sequence of discrete value samples $c_1, ..., c_N$. Our quantization approach starts by studying the pure behavior of the modulator as a mapping from the vector $\vec{x} = (x_1, x_2, ..., x_N) \in \mathbf{R}^N$ to the vector $\vec{c} = (c_1, c_2, ..., c_N) \in \mathbf{R}^N$. With this approach, the modulator obviously behaves like a vector quantizer since the output $\vec{c}$ only belongs to a discrete subset of $\mathbf{R}^N$. When the scalar quantizer has only a finite number of quantization levels, this set is even finite. In this paper, we assume that the scalar quantizer is uniform with step size q within its no-overload region (see Figure 2). As a consequence, the discrete output vector $\vec{c}$ necessarily belongs to a lattice of $\mathbf{R}^N$. Then, it can be immediately said that a $\Sigma\Delta$ modulator is a *lattice quantizer* in its own no-overload region, that is, the set of input vectors $\vec{x}$ such that Q does not overload. However the information yet to be determined is how the input vectors $\vec{x}$ are mapped to this lattice. This information is necessary to evaluate the performance of the vector quantizer.

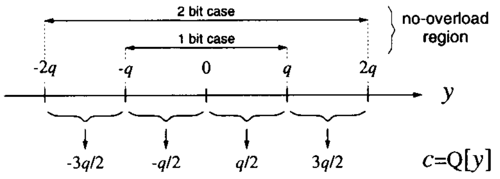

Figure 2: Transfer function of the scalar quantizer in its no-overload region.

0-7803-1865-X/94/$3.00 © 1994 IEEE

In our notations, this performance is measured by the mean squared error $\|\vec{c}-\vec{x}\|^2 = \sum_{k=1}^{N}|c_k - x_k|^2$ averaged over all input vectors $\vec{x}$ for a given probability distribution $\mathbf{p}$ in $\mathbf{R}^N$.

The mapping process of a vector quantizer as a many-to-one mapping, is usually presented as the succession of two separate steps, which are *encoding* and *decoding* [3]. The encoder is the part which decides what sets of input vectors will be given the same output codevector (since we have a many-to-one mapping). Thus the encoder defines a partition of the input space where each cell is the set of input vectors which will be assigned the same output codevector. The encoder will only provide an index which characterizes the partition cell into which the current input vector falls. Then, it is the role of the decoder to map each cell into a single vector representative, that is, the codevector. This decomposition normally appears by design. The performance of the vector quantizer is a result of the partitioning operation and the choice of codevectors. In $\Sigma\Delta$ modulation, this decomposition of course does not appear naturally since it was not designed as a vector quantizer. However, this structure can be extracted by block diagram transformation as shown in the next paragraph.

3 Encoding-decoding decomposition

In this paragraph, we show the encoder-decoder decomposition of a single-loop $\Sigma\Delta$ modulator assuming that the input vectors are restricted to the no-overload region. To perform the decomposition, we first transform the block diagram in order to get rid of the feedback loop. According to Figure 2, the transfer function of the scalar quantizer in the no-overload region can be expressed as $Q[y] = q\left(\left\lfloor \frac{y}{q}\right\rfloor + \frac{1}{2}\right)$, where $\lfloor z \rfloor$ designates the greatest integer smaller than or equal to z. The block diagram of the modulator with this explicit expression is shown in Figure 3(a). The equivalent block diagram of Figure 3(b) without feedback loop can be obtained thanks to the following equation, based on the notations of Figure 3(a):

$$\forall k \geq 1, \sum_{j=1}^{k}\left(\frac{c_j}{q}-\frac{1}{2}\right) = \left\lfloor \sum_{j=1}^{k}\left(\frac{x_j}{q}-\frac{1}{2}\right)\right\rfloor . \qquad (1)$$

This equation is proved in [4]. It shows the equality between a certain integrated version of the output and the quantized version of a certain integrated version of the input. Now, let us define

$$z_k = \sum_{j=1}^{k}\left(\frac{x_j}{q}-\frac{1}{2}\right) \quad \text{and} \quad a_k = \sum_{j=1}^{k}\left(\frac{c_j}{q}-\frac{1}{2}\right) \qquad (2)$$

with the convention $z_0 = a_0 = 0$. The sequence z_k can be obtained as shown by the block diagram of Figure 3(b). Thanks to equation (1), we have $a_k = \lfloor z_k \rfloor$ which is also implemented in the same block diagram. Finally (2) implies that $c_k = q\left(a_k - a_{k-1} + \frac{1}{2}\right)$ which is indeed the output of the block diagram. This shows the equivalence of Figures 3(a) and (b).

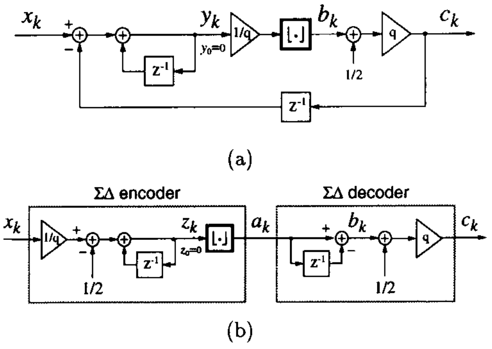

Figure 3: Equivalent block diagrams of the single-loop $\Sigma\Delta$ modulator in the no-overload region.

The encoding part is the front end portion of Figure 3(b) which makes the decision to merge whole sets of input signals into single values. It is therefore delimited by the quantization operator $\lfloor\cdot\rfloor$. One can check that the rest of the diagram only performs a one-to-one mapping transformation. The output of the encoder thus defined is an N-point sequence of integers $a_1, \ldots, a_N$. We directly take the N-tuple $A = (a_1, \ldots, a_N)$ as index to the partition cell corresponding to the output $a_1, \ldots, a_N$. The goal will be to determine the geometry of this partition. Then, the rest of the diagram can be interpreted as an operator which maps each possible index $A = (a_1, \ldots, a_N)$ into a codevector $\vec{c}$, and thus plays the role of the decoder. The goal will be to determine how each codevector is positioned by the encoder with respect to its corresponding cell.

The encoder which maps $\vec{x}$ into $A = (a_1, \ldots, a_N)$ can be considered as N subencoders working in parallel mapping $\vec{x}$ into a_k for $k = 1, \ldots, N$ respectively. Then, the partition defined by the whole encoder is the intersection of the partitions defined by these N subencoders respectively. Let us study the k^{th} subencoder. Since $a_k = \lfloor z_k \rfloor$, we would like to express z_k in terms of $\vec{x}$. Using the inner product notation $<\vec{x}, \vec{y}> = \frac{1}{N}\sum_{j=1}^{N} x_j y_j$, the expression of z_k in (2) can be written in the form $z_k = <\vec{d}_k, \vec{x}> - C_k$, where $\vec{d}_k = \frac{N}{q}\cdot(\overbrace{1, \ldots, 1}^{k \text{ times}}, 0, \ldots, 0)$ and $C_k = \frac{k}{2}$. Using these notations, the block diagram of the whole encoder can be equivalently represented as shown in Figure 4. It is easy to see that the partition defined by the k^{th} subencoder is composed of cells separated by hyperplanes of $\mathbf{R}^N$ perpendicular to $\vec{d}_k$ and equally spaced by $q_k = \frac{1}{\|\vec{d}_k\|}$. The vector $\vec{d}_k$ has the meaning of a wave vector. We therefore call this partition a *hyperplane wave partition*. Then the cells of the global partition are delimited by the hyperplanes of the N subpartitions altogether. Because the number of hyperplane directions is equal to the dimension of the space, the cells are simply N dimensional parallelepipeds. Their sides are respectively perpendicular to $\vec{d}_1, ..., \vec{d}_N$ and their vertices form a lattice in the no-overload region. One can see that the resulting partition is far from being a Voronoi partition.

It was shown in [5] that the codevector $\vec{c}$ corresponding to each parallelepipedic cell in the no-overload region is in fact the geometric center of the cell. This implies that, assuming a uniform probability distribution of the input vectors at least in the no-overload region, the $\Sigma\Delta$ decoder is indeed optimal with regard to the $\Sigma\Delta$ encoder.

4 Case of oversampled $\Sigma\Delta$ modulation

The previous paragraphs described the behavior of a $\Sigma\Delta$ modulator as a quantizer of non-constrained input sequences. In reality, $\Sigma\Delta$ modulation was designed to quantize the oversampled version of bandlimited continuous-time signals. We propose to study the situation where the bandlimited signals are T-periodic and uniformly sampled N times in the interval $]0, T]$. Such signals necessarily have the following finite discrete Fourier expansion

$$x(t) = X_1 + \sum_{i=1}^{p} X_{2i}\sqrt{2}\cos\left(2\pi i \tfrac{t}{T}\right) + X_{2i+1}\sqrt{2}\sin\left(2\pi i \tfrac{t}{T}\right), \tag{3}$$

where $X_1, X_2, ..., X_W$ are W real numbers with $W = 2p+1$. These continuous-time signals necessarily belong to a W dimensional space of signals. Then, the input sequences of the modulator have the form $x_k = x(\frac{k}{N}T)$ where $x(t)$ verifies (3). We assume that the oversampling ratio $R = \frac{N}{W}$ is greater than 1. It is easy to see that the corresponding vectors $\vec{x} = (x_1, ..., x_N)$ necessarily belong to a W dimensional subspace $\mathcal{V}$ of $\mathbf{R}^N$. Also, it can be easily shown that the norm of $\vec{x}$ is related to the MSE of $x(t)$ as $\|\vec{x}\|^2 = \frac{1}{T}\int_0^T |x(t)|^2\, dt$.

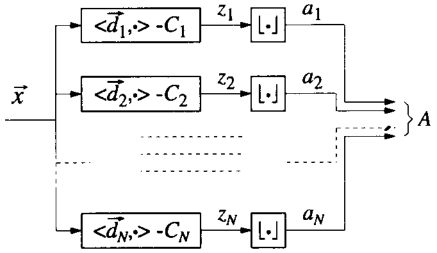

Figure 4: Equivalent structure of the $\Sigma\Delta$ encoder. The symbol $<\vec{d}_k, \cdot>$ designates the function which maps $\vec{x}$ into $<\vec{d}_k, \vec{x}>$.

Traditionally, an oversampled $\Sigma\Delta$ modulator is followed by a discrete-time lowpass filter to recover a bandlimited signal from the quantized output sequence $c_1, ..., c_N$. In our vector approach, this lowpass filtering amounts to an orthogonal projection of $\vec{c}$ onto the space $\mathcal{V}$. The global VQ diagram of oversampled $\Sigma\Delta$ modulation is shown in Figure 5. The system can be studied as a vector quantizer of the W dimensional space $\mathcal{V}$. Note that the MSE between the continuous-time versions $x(t)$ and $c'(t)$ of $\vec{x}$ and $\vec{c'}$ respectively, is equal to $\|\vec{c'} - \vec{x}\|^2$ which will be obtained through the study of the VQ performance.

The encoding part of the vector quantizer is in fact the same as in the previous general case. Its structure is still given by Figure 4. The difference is that the vectors $\vec{d}_k$ of this block diagram belong the space $\mathbf{R}^N$ which is larger than the space $\mathcal{V}$ of input vectors $\vec{x}$. This difference can be solved by introducing the vectors $\vec{d'_k}$ which are the orthogonal projections of $\vec{d}_k$ on $\mathcal{V}$ respectively. It is easy to see that $\forall \vec{x} \in \mathcal{V}$, $< \vec{d}_k, \vec{x} > = < \vec{d'_k}, \vec{x} >$. In the diagram of Figure 4 we can then replace $\vec{d}_k$ by $\vec{d'_k}$. As before, we conclude that an oversampled $\Sigma\Delta$ encoder is composed of N subencoders working in parallel which define N hyperplane wave partitions of $\mathcal{V}$ with wave vectors $\vec{d'_k}$ respectively. This time, the whole partition, still obtained from the intersection of these N subpartitions, is no longer composed of parallelepipedic cells, since the number N of hyperplane directions is superior to the dimension W of the input space $\mathcal{V}$. The resulting structure can be observed in Figure 6 which shows the partition defined by the modulator on the two dimensional space of T-periodic sinusoids with arbitrary phase and amplitude.

Now, one wonders whether the set of codevectors $\vec{c'}$ obtained by projection of the vectors $\vec{c}$ on $\mathcal{V}$ is still optimal with regard to the $\Sigma\Delta$ encoder in $\mathcal{V}$ for some input probability distribution. It was in fact shown in [5] that this set is not optimal and is not even consistent. This means that each possible codevector $\vec{c'}$ does not necessarily lies in its corresponding partition cell. Numerical tests performed in [5] also showed differences of performance between the linear decoder (using the bandlimitation filter) and a consistent decoder (having consistent codevectors). In the first case, the averaged MSE $\|\vec{c'} - \vec{x}\|^2$ was observed to decrease asymptotically with the oversampling ratio R in R^{-3}. This corresponds to the classical performance of the single-loop $\Sigma\Delta$ modulator [1]. In the second case, the asymptotic decrease was observed to be faster and in R^{-4}.

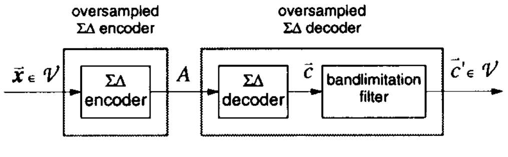

Figure 5: Vector quantization block diagram of oversampled $\Sigma\Delta$ modulation with linear decoding.

5 Intrinsic performance of oversampled $\Sigma\Delta$ modulation

Although the $\Sigma\Delta$ decoder part of Figure 3(b) is a built-in function of the $\Sigma\Delta$ modulator of Figure 1, we consider that the intrinsic performance of a $\Sigma\Delta$ modulator all depends on its encoding part. The reason is that after a signal is quantized by a $\Sigma\Delta$ modulator, the decoding choice can always (at least theoretically) be modified by postprocessing, since the mapping from the cell indices A to the codevectors $\vec{c}$ is invertible (see Figure 3(b)). Then, we define the intrinsic performance of the oversampled $\Sigma\Delta$ modulator as to be the VQ performance we would get if we used the oversampled $\Sigma\Delta$ decoder which is optimal with respect to the $\Sigma\Delta$ encoding part.

In this paragraph, we propose to derive a bound to this performance with the assumption of no-overloading input signals. Precisely, we derive a lower bound to the averaged MSE yielded by the optimal decoder for a given probability distribution p of $\vec{x}$ in the no-overload region, as a function of the oversampling ratio R. This MSE will be denoted by MSE_{opt}.

According to [6], such a lower bound can be derived if we know the number of cells M of the encoding partition in the input region. It was indeed shown that for M large enough,

$$MSE_{opt} \geq C(W, \mathbf{p}) \cdot M^{-2/W} \qquad (4)$$

where $C(W, \mathbf{p})$ is a coefficient which only depends on W and $\mathbf{p}$. For the derivation of an MSE lower bound, it is sufficient to derive an upper bound on M. This can be done thanks to the hyperplane wave structure of the encoding partition. It was indeed shown in [7] that

$$M \leq \binom{N}{W} (Dd+2)^W, \qquad (5)$$

where D is the diameter of the input region and d is the maximum length (or norm) of the wave vectors. Because orthogonal projections are non-expansive operators, we can write $\|\vec{d'_k}\|^2 \leq \|\vec{d_k}\|^2 = \frac{1}{N} \cdot \frac{N^2}{q^2} \cdot k$. This implies that $d \leq \frac{N}{q}$. It is easy to show that $\binom{N}{W} \leq \frac{N^W}{W!}$. Then, using (4), (5) and the last two inequalities, we have

$$MSE_{opt} \geq C(W, \mathbf{p}) \cdot W!^{2/W} N^{-2} \left(\frac{D}{q}N+2\right)^{-2}$$

Finally, using the relation $N = R \cdot W$, we obtain for R asymptotically large

$$MSE_{opt} \geq \left(\frac{q}{D}\right)^2 C'(W, \mathbf{p})\ R^{-4}$$

where $C'(W, \mathbf{p}) = C(W, \mathbf{p}) \cdot W!^{2/W} \cdot W^{-4}$. This proves that the asymptotic behavior of MSE_{opt} is lower bounded by $\mathcal{O}(R^{-4})$. This lower bound was shown in [8] in the case of constant inputs.

The numerical tests performed in [5] showed that there exists a decoder which achieves this asymptotic behavior. This evidence indicates that $\mathcal{O}(R^{-4})$ is not only a lower bound but might be achieved by MSE_{opt}.

References

[1] R.M.Gray, W.Chou, and P.-W.Wong, "Quantization noise in single-loop sigma-delta modulation with sinusoidal input," *IEEE Trans. Commun.*, vol. COM-37, pp. 956-968, Sept. 1989.

[2] O.Feely and L.O.Chua, "The effect of integrator leak in $\Sigma\Delta$ modulation," *Proc. IEEE Int. Symp. Circ. and Systems*, vol. CAS-38, pp. 1293–1305, Nov. 1991.

[3] A.Gersho and R.M.Gray, *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.

[4] N.T.Thao and M.Vetterli, "Lower bound on the mean squared error in multi-loop $\Sigma\Delta$ modulation with periodic bandlimited signals," *Proc. 27th Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA*, Nov. 1993.

[5] N.T.Thao and M.Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. on Signal Proc.* To appear in Mar. 1994.

[6] P.L.Zador, "Development and evaluation of procedures for quantizing multivariate distributions," *Ph.D. Dissertation, Stanford University*, 1963. Microfilm 64-9855.

[7] N.T.Thao and M.Vetterli, "Lower bound on the mean squared error in oversampled quantization of periodic signals," *IEEE Trans. Information Theory.* Submitted.

[8] A.Zakhor and K.Ibraham, "Lower bounds on the MSE of the single loop sigma delta modulator," *Proc. 23th Asilomar Conf. in Signals, Systems and Computers*, pp. 849-853, Oct. 1989.

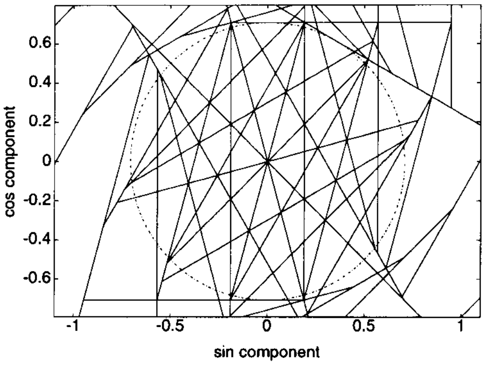

Figure 6: Partition defined by a single-bit single-loop $\Sigma\Delta$ modulator on the space of sinusoids of arbitrary phase and amplitude, sampled 12 times in one period. The circle represents the no-overload region.