Floating-Point Quantization Error Analysis in Subband Codec System

*Kyu-Sik Park and **Sung-Il Bang

Abstract

The very purpose of subband codec is the attainment of data rate compression through the use of quantizer and optimum bit allocation for each decimated signal. Yet the question of floating-point quantization effects in subband codec has received scant attention. There has been no direct focus on the analysis of quantization errors, nor on design with quantization errors embedded explicitly in the criterion.

This paper provides a rigorous theory for the modelling, analysis and optimum design of the general M-band subband codec in the presence of the floating-point quantization noise. The floating-point quantizers are embedded into the codec structure by its equivalent multiplicative noise model. We then decompose the analysis and synthesis subband filter banks of the codec into the polyphase form and construct an equivalent time-invariant structure to compute exact expression for the mean square quantization error in the reconstructed output. The optimum design criteria of the subband codec is given to the design of the analysis/synthesis filter bank and the floating-point quantizer to minimize the output mean square error. Specific optimum design examples are developed with two types of filter banks-orthonormal and biorthogonal filter bank, along with their perpormance analysis.

I. Introduction

The purpose of the subband coding is to decompose the input signal frequency band into a set of uncorrelated frequency bands by subband filtering and then to encode of these subbands using a bit allocation rationale matched to the signal energy in that subband. For the same distortion level, the total number of bits needed to transmit encoded subband signal is less than the number needed to transfer the signal directly. This reduction in bit rate is called (data) *compression*.

Subband coding has been proposed for many applications in the field of audio and video compression such as MPEGI, MPEGII, Dolby AC-3. Presently the modulated DCT type subband filter bank is the standard in those compression algorithms.

Over the last decade, the subband coding theory has been reached a high level of maturity of perfect reconstruction(PR) systems without considering coding errors such as quantization. In the absence of coding errors, many different classes of perfect reconstruction(PR) filter banks(FBs) have been described in the literature. Good accounts of the one-dimensional PR FB theory are given in [1]-[5] and the modulated DCT type filter bank is one of this kind. However, in the actual system, the signals are quantized before transmission to the receiver side and reconstructed by the synthesis FB. Thus the quantization effects must be carefully considered in the subband codec system design. In other words, to improve overall perfo-rmance of the subband codec by reducing the quantization effects, we must select the analysis/synthesis filter bank that minimize the quantization error in the codec.

In the presence of fixed-point quantizer in codec, westerlink, et. al[6] embed the equivalent noise model of the fixed-point quantizer in a two channel subband filter bank and analyze the resulting structure in a deterministic way. In Ref. [7], A. Tababai proposed the cyclo-stationary concept for the analysis of fixed point quantization effects in two channel filter banks. In Ref. [8], K. Park proposed the polyphase concept for the time-invariant analysis of quantization effects for general M-band subband codec. This approach avoids the complications arise from the time-varying analysis of the cyclo-stationary approach. However, all these papers are restricted to the fixed-point quantization.

Recently, N. Uzun[9] genratizes and extends the idea from the fixed-point case to the floating-point case in a two channel filter bank by using the time-varying nature (cyclostationary) of the subband signals. But the idea was restricted to the only two channel case because of the

*Dept. of Information and Telecommunication, Sangmyung University
**Dept. of Electronics Engineering, Dankook University

Manuscript Received: April 21, 1997.

analysis complexity.

The purpose of this paper is to provide rigorous analysis for the floating-point quantization effects in general M-band subband codec. We develop an optimum subband codec design methodology based on two concepts:

- design subband filter bank: we employ the polyphase decomposition technique and embed the multiplicative noise model for the floating-point quantizer. This polyphase construction and quantizer model ebnables us to calculate the MSE in a closed form time-domain formula. The minimization of this MSE is the objective of the optimum subband filter bank design described herein.
- design floating-point quantizer: the bit allocation that minimize the output MSE is the design criteria for the optmum quantizer.

Specific optimal design examples for two-channel orthonormal FB and biorthogonal FB are given to demonstrate our design methodology.

II. Background Theory

A. Floating-point quantization model

In floating-number number system, the number v is represented as

$$v = \text{sign}(v) m \beta^{e} \tag{1}$$

where sign(v) is the signum function, m is mantissa, e is exponents and β is the base of the floating-point number system. With a assumption of no under or overflows from the number range, there arise roundoff error only due to the rounding of mantissas, because exponents are integer. Roundoff error in floating-point can then be modeled as multiplicative error[10]. It is defined as

$$\epsilon = \frac{v' - v}{v} = \frac{(m' - m)}{m} \tag{2}$$

where v, v' are input to the quantizer and quantized output, m, m' are infinite precision mantissa and quantized mantissa, and ϵ is a mantissa roundoff error. Then the quantized output can be represented as

$$v' = vr = v(1 + \epsilon) = v + v\epsilon \tag{3}$$

where $r = 1 + \epsilon$ as shown in figure 1.

Note that the floating-point quantizer has the property that the roundoff error are zero mean, and orthogonal to

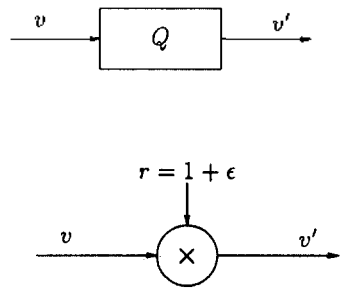


Figure 1. (a) Floating-point quantizer, (b) multiplicative noise model.

the input to the quantizer[10]

$$E[\epsilon] = 0, \quad E[\epsilon v] = 0 \tag{4}$$

B. Polyphase decomposition of M-band filter bank

The maximally decimated M-band filter bank structure with floating-point quantizers is shown in figure 2.

For this structure, we assume FIR filters of length NM for the analysis and synthesis filters. The total data rate in samples/second is unaltered from $x(n)$ to the set of subsampled signal $\{v_k(n),\ k=0, 1\cdots, M-1\}$ as implied by the term maximal decimation or critical subsampling.

Using the polyphase decomposition technique, we can express each analysis/synthesis filter $H_k(z)$, $G_k(z)$ in terms of M polyphase components

$$H_k(z) = \sum_{l=0}^{M-1} z^{-l} H_{k,l}(z^M), \tag{5}$$

$$G_k(z) = \sum_{l=0}^{M-1} z^{(M-1)-l} G_{k,(M-1-l)}(z^M).$$

Then we can represent the analysis filter bank in terms of the $M \times M$ polyphase matrix $\mathcal{H}_p(z)$ such as

$$\mathcal{H}_p(z) = \begin{bmatrix} H_{0,0}(z) & H_{0,1}(z) & \cdots & H_{0,M-1}(z) \\ H_{1,0}(z) & H_{1,1}(z) & \cdots & H_{1,M-1}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{M-1,0}(z) & H_{M-1,1}(z) & \cdots & H_{M-1,M-1}(z) \end{bmatrix} \tag{6}$$

where $H_{p,k}$ are kth polyphase components of the pth corresponding analysis filter. For the synthesis filter bank, we define a polyphase matrix $\mathcal{G}_p'(z) = J\mathcal{G}_p^T(z)$ in the same manner where J is the counter identity matrix[4]

$$J = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & & & & \vdots \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix} \tag{7}$$

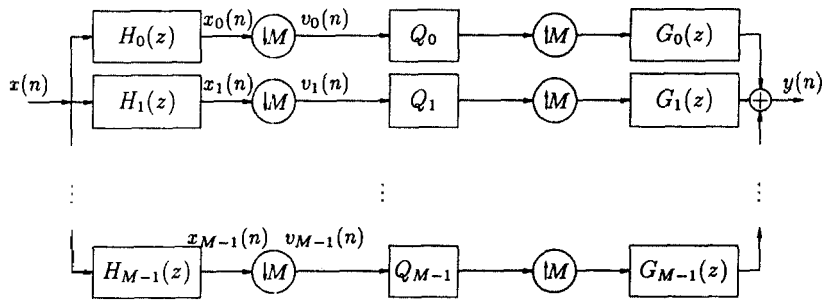


Figure 2. M-band subband codec system with floating-point quantizers

Then each multiplicative noise model for the floating-point quantizer is embedded into the structure. We replace the bank of filters by its polyphase equivalent and shift the samplers to the left and right of each polyphase matrix by using the noble identity shown in figure 3. This gives the polyphase equivalent structure of figure 4.

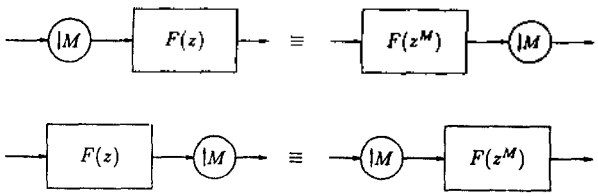


Figure 3. Noble identity.

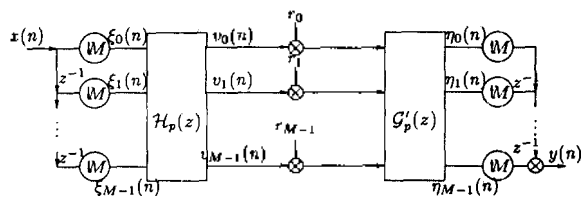


Figure 4. Polyphase equivalent structure

In our model of figure 4, v_i, v_i', ϵ_i are input to the quantizer, quantized output, and the quantization error for the ith channel. From Eq. (3), (4), we see that

$$E[r_i] = 1, \quad \sigma_{r_i}^2 = \sigma_{\epsilon_i}^2, \quad E[\epsilon_i \epsilon_j] = 0, \quad E[\epsilon_i v_i] = 0 \quad (8)$$

Figure 5 is a equivalent vector-matrix representation of figure 4 where $\underline{\xi}^T(n) = [\xi_0(n), \xi_1(n), \cdots, \xi_{M-1}(n)]$ and $\underline{v}(n)$, $\underline{r}(n)$, $\underline{\eta}(n)$ are similarly defined. At this point the system is time-invariant from $\underline{\xi}(n)$ to $\underline{\eta}(n)$ at the slow clock rate, $\frac{f_s}{M}$ where f_s is the sampling rate of the input signal.

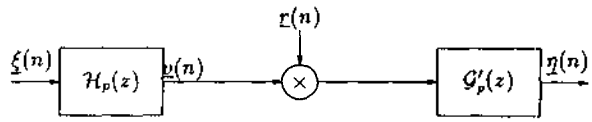


Figure 5. Vector-matrix equivalent structure

Without considering quantization from figure 5, the perfect reconstruction is achieved by satisfying the following sufficient condition

$$\mathcal{P}(z) = \mathcal{G}_p'(z)\mathcal{H}_p(z) = z^{-\mu} I_{M \times M} \quad (9)$$

so that the input signal $x(n)$ can be recovered from the reconstructed output $y(n)$ within a time shift.

Many different classes of FIR PR filter banks in the absence of quantizer errors have been described in the literature[1]-[5]. However, only two types of FIR filter bank, namely the orthonormal FB and the biorthogonal FB are considered in this paper.

Ⅲ. Quantization Error Analysis

We define the total quantization error as a difference

$$\underline{\eta}_q(z) = \underline{\eta}(z) - \underline{\eta}_o(z) \quad (10)$$

where the subscript "o" implies the system without the quantizers. From figure 4, the quantized output for the ith channel is

$$v_i'(n) = v_i(n) r_i(n) = v_i(n) + v_i(n)\epsilon_i(n) \quad (11)$$

By defining $v_{\epsilon_i}(n) \triangleq v_i(n)\epsilon_i(n)$, $V_{\epsilon_i}(z) = Z\{v_{\epsilon_i}(n)\}$, we have $V'(z) = V_i(z) + V_{\epsilon_i}(z)$ Then we see that the output vector from figure 5 becomes

$$\underline{\eta}(z) = \underline{\mathcal{G}}_p'(z)\underline{V}'(z) = \underline{\mathcal{G}}_p'(z)[\underline{V}(z) + \underline{V}_\epsilon(z)] \tag{12}$$

where $\underline{V}'(z) = [V_0'(z), V_1'(z), \cdots, V_{M-1}'(z)]^T$. For the case of no quantization, $\underline{\eta}_o(z)$ becomes

$$\underline{\eta}_o(z) = \underline{\mathcal{G}}_p'(z)\mathcal{H}_p(z)\underline{\xi}(z) = \underline{\mathcal{G}}_p'(z)\underline{V}(z) \tag{13}$$

where $\underline{V}(z) = \mathcal{H}_p(z)\underline{\xi}(z)$. Then the total quantization error at the reconstructed output due to the floating point quantizer is

$$\begin{aligned}\underline{\eta}_q(z) &= \underline{\eta}(z) - \underline{\eta}_o(z) \\ &= \underline{\mathcal{G}}_p'(z)[\underline{V}(z) + \underline{V}_\epsilon(z)] - \underline{\mathcal{G}}_p'(z)\underline{V}(z) \\ &= \underline{\mathcal{G}}_p'(z)\underline{V}_\epsilon(z)\end{aligned} \tag{14}$$

Since v and ϵ are uncorrelated from our quantization model, the output covariance matrice of $\underline{\eta}_q(z)$ can be easily derived as

$$R_{\underline{\eta}_q\underline{\eta}_q}[m] = \mathcal{G}_{p,-m}' * R_{\underline{v}_\epsilon\underline{v}_\epsilon}(m) * (\mathcal{G}_{p,m}'J)^T \tag{15}$$

Futhermore at $m=0$, it becomes

$$R_{\underline{\eta}_q\underline{\eta}_q}[0] = \sum_{j=0}^{N-1}\sum_{k=0}^{N-1}\{JG_{p,j}^T R_{\underline{v}_\epsilon\underline{v}_\epsilon}(j-k)G_{p,k}J\} \tag{16}$$

where $R_{\underline{v}_\epsilon\underline{v}_\epsilon}[m] = E[\underline{v}_\epsilon(n)\ \underline{v}_\epsilon^T(n+m)]$ is the $M \times M$ correlation matrix of $\underline{v}_\epsilon(n)$ and $G_{p,k}'$ is polyphase coefficient matrix of $\underline{\mathcal{G}}_p'(z) = \sum_{k=0}^{N-1} G_{p,k}' z^{-k}$.

We now consider the quantized system of figure 4, 5. We can demonstrate that $R_{\underline{\eta}\underline{\eta}}[0]$ is the covariance of the Mth block output vector

$$\begin{aligned}\underline{\eta}^T(n) &= [\eta_0(n), \eta_1(n), \cdots, \eta_{M-1}(n)] \\ &= [y(Mn), y(Mn+1), \cdots, y(Mn+M-1)]\end{aligned} \tag{17}$$

such that

$$R_{\underline{\eta}\underline{\eta}}[0] = \begin{bmatrix} R_{yy}(Mn, Mn) & \cdots & R_{yy}(Mn, Mn+M-1) \\ R_{yy}(Mn+1, Mn) & \cdots & R_{yy}(Mn+1, Mn+M-1) \\ \vdots & \ddots & \vdots \\ R_{yy}(Mn+M-1, Mn) & \cdots & R_{yy}(Mn+M-1, Mn+M-1) \end{bmatrix} \tag{18}$$

where $R_{yy}(Mn+k,\ Mn+j) = E[y(Mn+k)y(Mn+j)]$ for $k, j = 0, 1, \cdots, M-1$. Note that this is cyclostationary; the covariance matrix of the next block of M outputs will also equal to $R_{\underline{\eta}\underline{\eta}}[0]$. Each block of M output samples will thus have same sum of variances. We take the MS value of the output on the average of the diagonal elements of Eq. (18)

$$\begin{aligned}\sigma_y^2 &= \overline{E[y^2(n)]} \\ &= \frac{1}{M}\sum_{j=0}^{M-1} R_{yy}(Mn+j, Mn+j) = \frac{1}{M}\sum_{j=0}^{M-1} E[y^2(Mn+j)] \\ &= \frac{1}{M}\mathrm{Trace}\{R_{\underline{yy}}[0]\}.\end{aligned} \tag{19}$$

We now define quantization error at the system output as

$$y_q(n) \triangleq y(n) - y_o(n). \tag{20}$$

where $y(n)$ is the system output from figure 4 and $y_o(n)$ is the output without the quantizer. Then the total mean square(MS) quantization error at system output is

$$\sigma_{y_q}^2 \triangleq \overline{E[y_q^2(n)]} = \frac{1}{M}\mathrm{Trace}[R_{\underline{\eta}_q\underline{\eta}_q}[0]]. \tag{21}$$

Furthermore, by substituting Eq. (16) into Eq. (21), we obtain

$$\sigma_{y_q}^2 = \frac{1}{M}Trace[\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} G_{p,j}^T R_{\underline{v}_\epsilon\underline{v}_\epsilon}(j-k)G_{p,k}] \tag{22}$$

Then by expanding the polyphase coefficient matrices $\{G_{p,j},\ j=0, 1, \cdots, M-1\}$ in terms of the corresponding synthesis filter coefficients

$$G_{p,j} = \begin{bmatrix} g_0(Mj) & g_0(Mj+1) & \cdots & g_0(Mj+M-1) \\ g_1(Mj) & g_1(Mj+1) & \cdots & g_1(Mj+M-1) \\ \vdots & \vdots & \ddots & \vdots \\ g_{M-1}(Mj) & g_{M-1}(Mj+1) & \cdots & g_{M-1}(Mj+M-1) \end{bmatrix} \tag{23}$$

in Eq. (22), and by using

$$R_{v_\epsilon v_\epsilon}[m] = R_{v_i v_j}[m]R_{\epsilon_i\epsilon_j}[m],\ R_{\epsilon_i\epsilon_j}[m] = \sigma_{\epsilon_i}^2\delta_{i-j}\delta[m] \tag{24}$$

where $v(n)$ and $\epsilon(n)$ uncorrelated, we can show that the total MS quantizatin error at the output is

$$\sigma_{y_q}^2 = \frac{1}{M}\sum_{i=0}^{M-1}\sigma_{v_i}^2\sigma_{\epsilon_i}^2\sum_{l=0}^{MN-1} g_i^2(l) \tag{25}$$

where $\sigma_{v_i}^2$ is the variance of subband signal and $\sigma_{\epsilon_i}^2$ is the variance of quantization error on the ith channel of the subband codec.

From Eq. (25), the variance of the subband signal, $\sigma_{v_i}^2$,

can be obtained from the correlation function $R_{v_i v_i}(m)$ such that

$$\sigma_{v_i}^2 = R_{v_i v_i}(0) = R_{x_i x_i}(0) = \sum_k \sum_l h_i(k) h_i(l) R_{xx}(k-l) \tag{26}$$

To get the variance of quantization error, $\sigma_{\epsilon_i}^2 = var[\epsilon_i]$, we assume that the input $x(n)$ is AR(1) source with unit variance such that

$$R_{xx}(m) = \sigma_x^2 \rho^{|m|} \tag{27}$$

where $\sigma_x^2 = var(x) = 1$ and ρ is correlation coefficient. With an floating-point number in a form of $v = m2^e$ where m is mantissa between 1/2 and 1, and e is an integer exponent, if we quantize mantissa to R bits, then the quantized output

$$v' = vr = v(1+\epsilon), \quad -2^{-R} \le \epsilon \le 2^{-R} \tag{28}$$

where ϵ is assumed to be uniformly distributed in the range $(-2^{-R}, 2^{-R})$. This assumption is valid when the word length is not too short[10]. In this case, the noise variance is given by

$$\sigma_\epsilon^2 = \frac{1}{12} 2^{-2R} \tag{29}$$

Finally, by taking account Eq. (26) and (29) into Eq. (25) the output MS quantization error becomes

$$\begin{aligned} \sigma_{y_q}^2 &= \frac{1}{M} \sum_{i=0}^{M-1} \sigma_{\epsilon_i}^2 \sigma_{v_i}^2 \sum_{l=0}^{MN-1} g_i^2(l) \\ &= \frac{1}{M} \sum_{i=0}^{M-1} [\sum_k \sum_l h_i(k) h_i(l) R_{xx}(k-l)] [\frac{1}{12} 2^{-2R_i}] \sum_{l=0}^{MN-1} g_i^2(l) \end{aligned} \tag{30}$$

Thus we have formulated the output MSE in terms of the analysis/synthesis filter coefficients $h_i(n)$, $g_i(n)$ of length MN, input autocorrelation function $R_{xx}[m]$, and R_i the bits allocated to each channel.

Ⅳ. The Optimum Filter Banks

In this section, we briefly review the properties of orthonormal and biorthogonal FB. The PR conditions in the absence of quantization error is given for the two-channel filter banks. These type of filter banks will be used as a optimum design examples in the next section.

Orthonormal filter Bank

Perfect reconstruction(PR) can be achieved by orthonormal(or lossless) filter bank[1][5]. Orthonormality condition implies

$$\tilde{\mathcal{H}}_p(z) \mathcal{H}_p(z) = I_{M \times M}, \quad \tilde{\mathcal{H}}_p(z) \triangleq \mathcal{H}_p^T(z^{-1}). \tag{31}$$

This in turn means that the synthesis polyphase matrix is also lossless from the sufficient PR condition Eq. (9).

In the time domain, this condition is shown to be

$$\begin{aligned} \sum_k h_r(k) h_s(Mn+k) &= \delta_{(r-s)} \delta(n), \\ \sum_k g_r(k) g_s(Mn+k) &= \delta_{(r-s)} \delta(n). \end{aligned} \tag{32}$$

We now recall the MSE equation (30) and note the orthonormality condition in Eq. (32). Then the MS quantization error reduces to a rather simple form

$$\sigma_{y_q}^2 = \frac{1}{M} \sum_{i=0}^{M-1} \sigma_{v_i}^2 \frac{1}{12} 2^{-2R_i} \tag{33}$$

since $\sum_l g_i^2(l) = 1$.

For the two-band(M = 2) orthonormal FB, we select $H_0(z)$ an N-tap FIR filter(N = even) and then choose

$$\begin{aligned} H_1(z) &= z^{-(N-1)} H_0(-z^{-1}) \Longleftrightarrow h_1(n) \\ &= (-1)^{n+1} h_0(N-1-n) \end{aligned} \tag{34}$$

to satisfy orthonormal requirements. Then the synthesis filters are given by

$$\begin{aligned} G_0(z) &= z^{-(N-1)} H_0(z^{-1}) \iff g_0(n) = h_0(N-1-n), \\ G_1(z) &= z^{-(N-1)} H_1(z^{-1}) \iff g_1(n) = h_1(N-1-n) \\ &= (-1)^n h_0(n). \end{aligned} \tag{35}$$

B. Biorthogonal filter Bank

The two-band biorthogonal FB[2][4] is a generalization of the orthonormal FB. This structure permits linear phase FIR PR filters in the two-channel case-a feature not possible with orthonormal constraints. Although unequal length biorthogonal PR FB is possible, we consider only equal length case for comparison with the orthonormal structure.

For the equal length analysis/synthesis FB, we consider the causal analysis filters, symmetric $H_0(z)$ and antisymmetric $H_1(z)$ of equal length L = even. Then the perfect reconstruction can be achieved by choosing the synthesis filters by

$$\begin{aligned} G_0(z) &= H_1(-z) \iff g_0(n) = (-1)^n h_1(n), \\ G_1(z) &= -H_0(-z) \iff g_1(n) = (-1)^{(n+1)} h_0(n). \end{aligned} \tag{36}$$

Then Eq. (36) implies the equal length analysis/synthesis filters and the causal synthesis filters with a symmetric $G_0(z)$ and a antisymmetric $G_1(z)$.

The biorthogonal perfect reconstruction is satisfied by

$$\begin{aligned} \sum_k h_i(k)\tilde{g}_i(k-(2n+1)) &= \delta(n-n_0) \\ \sum_k h_i(k)\tilde{g}_j(k-(2n+1)) &= 0 \quad \text{for } i \neq j \end{aligned} \tag{37}$$

where $\tilde{g}_j(n) \triangleq g_j(-n)$ and n_0 is some delay. We note that no simplification in MSE Eq. (30) is possible with biorthogonal FB case and the MS quantization error equation remains same as Eq. (30).

Our design problem is now to find the optimal PR filter bank and bits allocated to each channel which minimizes Eq. (33) for the orhonormal FB structure and Eq. (30) for the Biorthogonal FB structure.

V. Design Example and Performance Analysis

In this section, we have developed specific design examples for two different classes of filter banks, the orthonormal and biorthogonal two-channel case with equal length 6 tap filter banks. Our design problem is now to find the optimal PR filter bank and bit allocation which minimizes the output MSE for a given total bit allocation. We assume that each quantizer takes only integer bits and the high frequency components of the subband signal gets at least 1 bit. Otherwise there is no way to recover high frequency component of input signal at the output.

Our optimization algorithm tests for all possible bit combinations for the given average bit rate R bits/sample, calculates the optimal filter coefficients and MSE. It chooses the one with the minimum MSE among them. This is implemented by using IMSL FORTRAN Library (DNCONF). This package solves a general nonlinear constrained minimization problem using the successive quadratic programming algorithm and a finite difference gradient.

A. Orthonormal filter bank

The simulation results for the orthonormal FB are shown in Table 1, 2 for the input correlation $\rho = 0.95$, 0.75, 0.55.

Tabl2 1 lists the optimum integer bits allocated to each channel R_0, R_1 and the calculations of the output MSE based on Eq. (33). This table shows that as the average bit rate R bit/sample and input correlation ρ gets larger, the output MSE is getting smaller.

Table 1. MSE for orthonormal FB at $\rho = 0.95$, 0.75, 0.55 with respect to R

R	R_0	R_1	$MSE(\rho=0.95)$	$MSE(\rho=0.75)$	$MSE(\rho=0.55)$
1	1	1	0.020833	0.031812	0.038332
1.5	2	1	0.005489	0.006716	0.008162
2	3	1	0.001646	0.003187	0.004995
2.5	4	1	0.000687	0.002305	0.004202
3	5	1	0.000447	0.002084	0.004004

Table 2. Optimum orthonormal filter coefficients $h_0(n)$ at $\rho = 0.95$

R	$h_0(0)$	$h_0(1)$	$h_0(2)$	$h_0(3)$	$h_0(4)$	$h_0(5)$
1	0.31379	0.67552	0.60153	−0.07663	−0.23088	0.11326
1.5	0.38532	0.74741	0.48268	−0.08832	−0.14452	0.05135
2	0.38568	0.79648	0.42752	−0.14086	−0.10669	0.05168
2.5	0.38567	0.79628	0.42813	−0.14085	−0.10669	0.05167
3	0.38585	0.79633	0.42787	−0.14089	−0.10661	0.05166

The corresponding optimal filter coefficients of $h_0(n)$ for $\rho = 0.95$ are shown in Table 2. As seen from the tables, the optimal filter coefficients are quite insensitive to changes in average bit rate R (and also to ρ) although the output MSE is highly dependent on them.

The magnitude reponse of analysis filters $H_0(z)$, $H_1(z)$ corresponding to the designed filter coefficents of $\rho = 0.95$, 0.75 are shown in figure 6(a), (b) respectively for the given average bit rate $R = 3$.

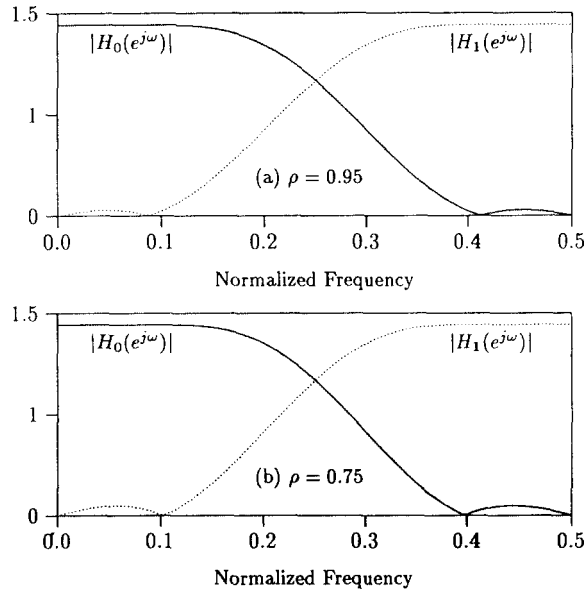


Figure 6. Magnitude response $H_0(z)$, $H_1(z)$ of orthonormal FB at $R = 3$.

From the figures, we clearly see the effect of input correlation change in magnitude responses. As the input correlation ρ decreases, the stopband ripples of each filter and the spillover from one band to another are getting larger. Consequently, more aliasing is introduced between the channels which yields larger MSE at the output. As expected, the magnitude reponses show the mirror image property at $\omega = \pi/2$.

B. Biorthogonal filter bank

The optimal designs for the equal length analysis/synthesis biorthogonal FB based on the minimization of Eq. (30) are shown in Table 3 and 4 for the case of input correlation $\rho = 0.95, 0.75, 0.55$.

Table 3. MSE for biorthogonal FB at $\rho = 0.95, 0.75, 0.55$ with respect to R

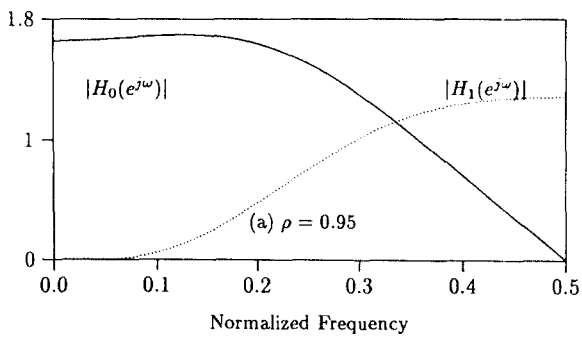
R	R_0	R_1	$MSE(\rho=0.95)$	$MSE(\rho=0.75)$	R_0	R_1	$MSE(\rho=0.55)$
1	1	1	0.012717	0.016786	1	1	0.038332
1.5	2	1	0.003763	0.006090	2	1	0.008162
2	3	1	0.001331	0.003136	3	1	0.004995
2.5	4	1	0.000643	0.001522	3	2	0.001999*
3	5	1	0.000334	0.000784	4	2	0.001270*

Table 4. Optimum biorthogonal filter coefficients $h_0(n)$, $h_1(n)$ at $\rho = 0.95$

R	$h_0(0)$	$h_0(1)$	$h_0(2)$	$h_1(0)$	$h_1(1)$	$h_1(2)$
1	0.25137	−0.68345	0.92910	0.08586	−0.23346	0.68665
1.5	0.09364	−0.46631	0.94227	0.04160	−0.20717	0.62902
2	−0.00552	−0.26864	0.93553	−0.00345	−0.16801	0.58269
2.5	−0.03304	−0.13232	0.92970	−0.00328	−0.13165	0.55537
3	−0.03332	−0.07694	0.92863	−0.00495	−0.11440	0.54612

From the tables, we see that the optimal filter coefficients are very sensitive to changes both in the average bit rate R and in the input correlation ρ.

Figure 7 shows the magnitude frequency responses corresponds to input correlation $\rho = 0.95$, 0.75 and $R = 3$.



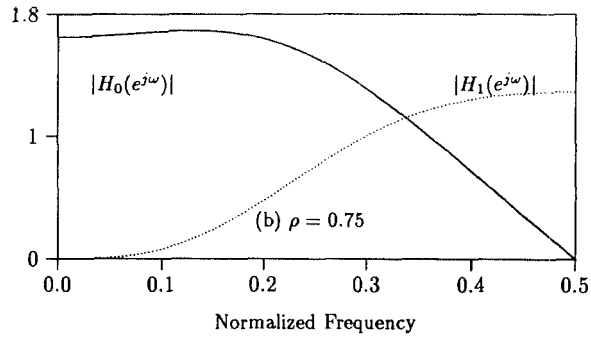


Figure 7. Magnitude response $H_0(z)$, $H_1(z)$ of biorthogonal FB at $R = 3$

We observe the significant overlaps in the stopband and the spillover between $H_0(z)$ and $H_1(z)$ as the input correlation decreases. These cause the increasement of the output MSE. Unlike the orthonormal case, the low and high-pass filters of biorthogonal analysis FB are not mirrors of each other, but they are linear phase.

C. Performance comparison

Optimum orthonormal and biorthogonal FB structures for the two channel case have been developed and demonstrated with AR(1) gaussian input signal with different input correlations.

Table 1 and 3 demonstrate that the biorthogonal FB is superior to the orthonormal counterpart in terms of output MS quantization error. Figure 8 compares the performance of orthonormal FB and biorthogonal FB in terms of output MSE.

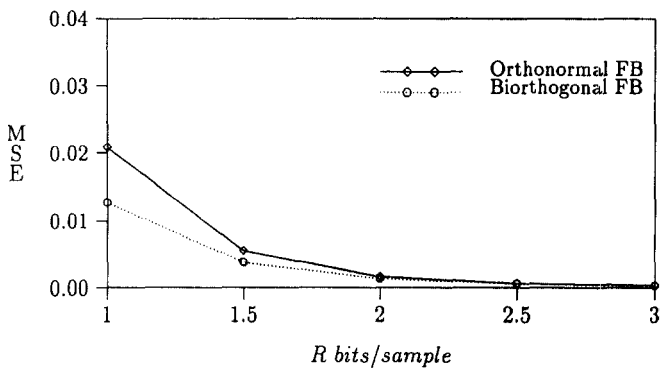


Figure 8. MSE comparison for orthonormal and biorthogonal FB for $\rho = 0.95$.

Table 2 and 4 shows that the orthonormal filter coefficients are insensitive to changes in input signal statistics. Hence the orthonormal FB is robust. On the other hand, the biorthogonal filter is very sensitive to these

parameters which in turn suggest a possible adaptive biorthogonal structure.

Ⅵ. Summary and Conclusions

We have presented a methodology for modelling, analysis for the floating-point quantization effects in general M band subband codec. This approach also sets up a quantization error measure which is to be minimized to design the optimum subband codec: subband filter bank and the quantizer.

Optimum orthonormal and biorthogonal FB structure have been designed for the 2 channel case. For the 2 channel equal-length filters, it turns out that the biorthogonal FB is superior to the orthonormal FB in terms of output MSE. However the orthonormal FB provides a robust system to changes in input signal statistics while the biorthogonal FB is very sensitive to these parameters. This suggests an adaptive biorthogonal structure that revises the bit allocation based on the measured changes in input correlation ρ. These changes are transmitted to the receiver as side information. We note that similar arguements were made for fixed-point quantization case in Ref. [8].

References

1. P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial," *Proc. IEEE*, vol. 78, pp. 56-93, Jan. 1990.
2. M. Vetterli and D. LeGall, "Perfect reconstruction FIR filter banks: Some properties and factorization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1057-1071, July 1989.
3. R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
4. A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*. Academic Press, 1992.
5. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
6. P. H. Westerink, J. Biemond, D. E. Boekee, "Scalar quantization error analysis for image subband coding using QMF's," *IEEE Trans. Signal Processing*, vol. 40, no. 2, pp. 421-428, Feb. 1992.
7. Ali Tabatabai, "Optimum Analysis/Synthesis Filter Bank Structures with Application to Sub-Band Coding Systems," pp. 823-826, *IEEE ISCAS88*, 1988.
8. Kyusik Park and R. A. Haddad, "Modelling, Analysis, and Optimum Design of Quantized M-band Filter Banks," *IEEE Trans. on Signal Processing*, Nov. 1995.
9. "Cyclostionary Modelling, Analysis, and Optimal Compensation of Quantization Effects in Subband Codec," *Ph. D Thesis*, Polytechnic University, Brooklyn, NY, 1993.
10. Jarmo Kontro, Kari Kalliojarvi and Yrjo Neuvo, "Floating-Point Arithmetic in Signal Processing," IEEE ISCAS '92, pp. 1784-1787, San Diego CA, 1992.

▲Kyu-Sik Park

Kyusik Park received B.S, M.S, and Ph.D degrees in 1986, and 1994, respectively, all from the Department of Electrical Engineering of Polytechnic University, Brooklyn, NY, USA.

In 1994, he joined the Semiconductor Division of Samsung Electronics as a staff engineer. He is currently a full-time lecturer with the Department of Information and Telecommunication in Sangmyung University, Chungchongnam-Do, Korea. His research interests are digital signal and image processing, and multimedia communication.

▲Sung-Il Bang

Sung-Il Bang received B.S, M.S, and Ph.D degrees in 1984, 1986, and 1992, respectively, from the Department of Electrical Engineering of Dankook University, Seoul, Korea.

In 1992, he served as a research director in Daegi Telecommunication, Ltd. He is currently and assistant professor with the Department of Electrical Engineering in Dankook University, Seoul, Korea. His research interests are digital signal processing, digital communication and mobile communication system.