Statistical Design of Analysis/Synthesis Systems with Quantization

AMIR DEMBO, MEMBER, IEEE, AND DAVID MALAH, FELLOW, IEEE

Abstract—A statistical model is used for the optimal design of analysis/synthesis systems which include quantization of the signals in separate bands. Two types of quantization approaches are analyzed: fine quantization modeled by additive noise, and matrix quantization (which includes vertical and horizontal vector quantizations and scalar quantization with fixed bit allocation as special cases). With proper modification, the general framework is also applicable for other types of quantization approaches. Based on the assumption that the system operates as a waveform coder, two error measures are used. The first is a generalization of the usual statistical mean-square error (MSE) to time-varying systems, whereas the second involves the MSE between the outputs of the analysis of the original and reconstructed signals. Both error measures are shown to be equivalent. We present a design method which is applicable for the design of optimal synthesis filters, given the analysis filters. In addition, for fine quantization, an iterative algorithm for the design of an optimal analysis/synthesis system is presented, together with its convergence properties. It is further shown that if no quantization is applied, the results obtained with the new method coincide with previously reported results.

I. Introduction

ANALYSIS/SYNTHESIS (A/S) systems are widely used in speech processing [1]–[8]. A typical application is medium rate waveform coding (e.g., [2]) where quantization is applied to the signals in separate frequency bands.

Known methods for the design of analysis/synthesis windows (interpreted also as analysis/synthesis filters in [1]) are based on deterministic error measures [4]–[6], and they ignore signal quantization. Thus, in the presence of quantization, even a unity A/S system is not necessarily an optimal one, and its performance depends also on the input statistics and quantization characteristics.

A statistical approach which takes into account the quantization effect in the design process is therefore needed. Such an approach is presented here for two common types of quantization. The first is "fine quantization" which is used in waveform coding schemes at a 16 kbit/s rate and above. This type of quantization is reasonably modeled by additive noise which is independent of the input signal. The second is vector/matrix quantization with a small (typically 256–1024) overall number of different codewords used to represent blocks of consecutive output vectors of the analysis stage. The discussion is restricted to quantization schemes aimed to work as waveform coders. These schemes have been recently used in conjunction with subband coding (cf. [7]). Scalar quantization with typically 1–4 bits per frequency band (e.g., [8]) is also regarded here as a particular case of vector/matrix quantization (single element). A/S systems are usually used in conjunction with the DFT transform. However, the statistical approach is applied here to a more general model in which any linear regular transform may be used.

The weighted overlap-add (WOLA) synthesis method [9], [10] is used throughout. In the next section, we review the A/S systems with WOLA synthesis, and we introduce the statistical models of the various quantization approaches. Since A/S systems are in general time varying, a generalization of the standard waveform error measures to a class of nonstationary random processes is needed and is presented in Section III. It is then used as an optimality criterion. Furthermore, the generalized error measure which is defined in the time domain and an error measure defined in the transform domain are shown to be equivalent. The details of the derivations are given in Appendix A. In Section IV, the optimal synthesis filters for given analysis and quantization subsystems are derived as the solution of a set of linear equations. Conditions for uniqueness of the solution, as well as some examples for which closed-form solutions exist, are also presented there. Certain asymptotic properties of the solution (as the quantization noise approaches zero) and its relation to results obtained with known design methods [5], [6] are discussed. An analysis of the computational complexity of the proposed design process is also given. Proofs and details of the subjects discussed in Section IV are given in Appendix B.

In the case of fine quantization (and only in this case), the dependence of the error measure on the analysis window is given explicitly by a positive semidefinite (PSD) quadratic form. Thus, similarly to [4], we present in Section V an iterative algorithm for the design of optimal A/S systems with fine quantization. This algorithm partially extends the results of [4] by incorporating the quantization and input signal characteristics into the design process. Furthermore, we present here its convergence properties which were not reported in [4], with details given in Appendix C. Conclusions are drawn in the last section.

Manuscript received April 25, 1986; revised September 16, 1986.

A. Dembo was with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. He is now with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

D. Malah is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 8718615.

0096-3518/88/0300-0328$01.00 © 1988 IEEE

II. Statistical Model of the A/S System

We present below a statistical model of a fixed time reference [1], [9] A/S system with WOLA synthesis and quantization of the output of the analysis stage. Although most of the A/S systems contain the DFT transform in the analysis stage, so that the quantizer input is the discrete short-time Fourier transform (DSTFT) of the input signal [10], some contain other linear regular transforms (such as DCT or Hadamard). Therefore, the modeling of an A/S system with an arbitrary linear regular transform is presented in order to apply the new design method to various coding systems. A schematic description of the A/S system is given in Fig. 1. In order to simplify the presentation of the new design method, we frequently use infinite sums so as not to bother with explicitly stating their limits, although they are actually finite—unless explicitly stated. The time domain sequences as well as the analysis and synthesis windows have real values, whereas the transform values may be complex. A detailed description of the model is given below.

1) The input signal sequence $x(\cdot)$ is multiplied by the sliding analysis window $h(\cdot)$ having a length of L_h samples. A time aliasing operation reduces these L_h values into a vector of length M (the transform size), denoted by $\boldsymbol{x}_s$, where the sth sample of $x(\cdot)$ is the one multiplied by $h(0)$. These vectors (for different time instances) are decimated with a decimation factor R, $1 \le R \le M$. The mth element of the vector $\boldsymbol{x}_{sR}$ is given by

$$x_{sR}(m) = \sum_{r=-\infty}^{\infty} h(sR - m - Mr)\, x(m + Mr), \quad 0 \le m \le M - 1. \tag{1}$$

A linear regular transform of size M operates on these vectors and results in output vectors $\boldsymbol{X}_{sR}$ of M elements each. This transform is represented by the $M \times M$ matrix T whose (k, m) element is denoted by $t(k, m)$, $0 \le k, m \le M - 1$. The kth element of the vector $\boldsymbol{X}_{sR}$ is given by

$$X_{sR}(k) = \sum_{m=0}^{M-1} t(k, m)\, x_{sR}(m), \quad 0 \le k \le M - 1. \tag{2}$$

If the DFT transform is applied [i.e., $t(k, m) = \exp(-j(2\pi km/M))$], the sequence of vectors $\boldsymbol{X}_{sR}$ is the DSTFT of the input signal. For the general transform represented by T, we denote the sequence of vectors $\boldsymbol{X}_{sR}$ as the discrete short-time transform (DSTT) of the input signal. This completes the analysis part of the system.

2) Let the quantized DSTT sequence of vectors be denoted by $\hat{\boldsymbol{X}}_{sR}$. Since the signal is changed by the quantization, it can be considered as a *modified* DSTT which we denote as MDSTT. Before describing in detail the various quantizers considered [i.e., the various types of mappings of $\hat{\boldsymbol{X}}_{sR}(\boldsymbol{X}_{sR})$], we describe the WOLA synthesis scheme which is used in all the systems considered here.

An inverse transform operates on the MDSTT and results in time domain vectors $\hat{\boldsymbol{x}}_{sR}$ of M elements each. The

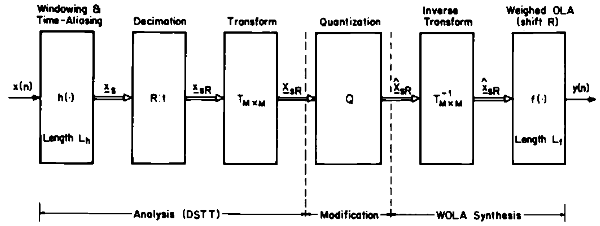


Fig. 1. Block diagram of an analysis/synthesis system with an arbitrary regular linear transform, modification (quantization), and WOLA synthesis.

inverse transform is represented by the matrix T^{-1} of dimensions $M \times M$, whose (m, k) element is denoted by $t^{-1}(m, k)$ for $0 \le m, k \le (M - 1)$. The mth element of $\hat{\boldsymbol{x}}_{sR}$ is thus given by

$$\hat{x}_{sR}(m) = \sum_{k=0}^{M-1} t^{-1}(m, k)\, \hat{X}_{sR}(k), \quad 0 \le m \le M - 1. \tag{3}$$

Each of the time domain vectors $\hat{\boldsymbol{x}}_{sR}$ is periodically extended, and a weighted overlap-add operation (with shift R) reconstructs the output sequence $y(n)$ [10, p. 321]. This operation is done using a synthesis filter $f(\cdot)$ of L_f samples, and is described by

$$y(n) = \sum_{s=-\infty}^{\infty} f(n - sR)\, \hat{x}_{sR}((n)_M) \tag{4}$$

where $(n)_M$ denotes $n(\text{mod } M)$.

3) Two types of quantization approaches are considered.

a) Fine Quantization (FQ): With fine quantization, typically more than 4 bits per input sample are used, and the output signal-to-noise ratio (SNR) is quite high. It is well known that for this case, the quantization error can be reasonably modeled as an additive noise vector that is uncorrelated with the input signal [11]. We shall further assume that these noise vectors are samples of a wide-sense stationary process with zero mean and known covariance sequence. Let $\boldsymbol{V}_{sR}$ denote the noise vector added to $\boldsymbol{X}_{sR}$; then

$$\hat{X}_{sR}(k) = V_{sR}(k) + X_{sR}(k), \quad 0 \le k \le M - 1. \tag{5}$$

For covenience of the presentation, we denote by $\boldsymbol{v}_{sR}$ the inverse transform of the vector $\boldsymbol{V}_{sR}$ and use the covariance sequence of $\boldsymbol{v}_{sR}$ rather than the covariance sequence of $\boldsymbol{V}_{sR}$. Thus,

$$\hat{x}_{sR}(m) = x_{sR}(m) + v_{sR}(m), \quad 0 \le m \le M - 1 \tag{6}$$

and

$$E[v_{sR}(m)] = E[v_{sR}(m)\, x(n)] = 0 \quad \forall n, s;\ 0 \le m \le M - 1 \tag{7a}$$

$$E[v_{sR}(m)\, v_{(s+d)R}(n)] = \Psi_{m,n}(dR) \quad \forall s, d;\ 0 \le m, n \le M-1. \tag{7b}$$

For example, if the vectors V_{sR} are uncorrelated both in time and in the transform domain, then $\Psi_{m,n}(dR) = 0$ for $d \neq 0$. Furthermore, if the DFT transform is used, the matrix whose elements are $\Psi_{m,n}(0)$ is a circulant matrix whose first row is the IDFT of dimension M of the sequence $(E[V_{sR}(0)^2], \cdots, E[V_{sR}(M-1)^2])$. The latter sequence is closely related to the bit allocation used in the different bands.

b) Matrix Quantization (MQ): The matrix quantization of blocks of $B \ge 1$ DSTT vectors is explained below.

The vectors $\{X_{sBR}, \cdots, X_{(sB+(B-1))R}\}$ are regarded as an $M \times B$ matrix $\mathbf{X}_{s(BR)}$. The codebook contains L different matrices $\{C^{(1)}, \cdots, C^{(L)}\}$ and the space of all $M \times B$ complex matrices is divided into L distinct sets $\{A^{(1)}, \cdots, A^{(L)}\}$ such that $\cup_{i=1}^{L} A^{(i)} = {}^{M \times B}$. At the time instance s, the coder selects the index i_s, $1 \le i_s \le L$ according to

$$i_s = l \quad \text{if and only if} \quad \mathbf{X}_{s(BR)} \in A^{(l)} \quad 1 \le l \le L. \tag{8a}$$

When a specific index value l is received by the decoder, it uses the matrix $C^{(l)}$ as $\hat{\mathbf{X}}_{s(BR)}$, and therefore the sequence $\{c_0^{(l)}, \cdots, c_{B-1}^{(l)}\} = T^{-1}C^{(l)}$ is used as the sequence of vectors $\{\hat{x}_{sBR}, \cdots, \hat{x}_{(sB+(B-1))R}\}$ in the synthesis. This sequence is assumed to be real even for complex transforms (one can easily guarantee this property by an appropriate selection of the codebook matrices). The output samples of the A/S system are the result of a WOLA synthesis [using (4)], applied to the vectors:

$$\hat{x}_{(sB+d)R} = c_d^{(i_s)}, \quad 0 \le d \le B-1, \quad 1 \le i_s \le L, \quad -\infty < s < \infty. \tag{8b}$$

It is quite difficult to design a complete codebook for large values of L (e.g., 1024 and above), and therefore the Cartesian product of several codebooks is usually used. In such a scheme, the index i_s is essentially a vector of length $J \ge 1$, i.e., $i_s = (i_s(1), \cdots, i_s(J))$ where $1 \le i_s(k) \le L_k$ with $L = \Pi_{k=1}^{J} L_k$. The reduction in complexity is obtained by designing J *independent* codebooks, with the kth codebook being described by the pair of sets $\{A^{(1,k)}, \cdots, A^{(L_k,k)}\}$, $\{C^{(1,k)}, \cdots, C^{(L_k,k)}\}$, and the coding–decoding scheme is

$$i_s(k) = l \quad \text{if and only if} \quad \mathbf{X}_{s(BR)} \in A^{(l,k)}, \quad 1 \le l \le L_k,\ 1 \le k \le J \tag{9a}$$

$$\hat{x}_{(sB+d)R} = \sum_{k=1}^{J} c_d^{(i_s(k),k)}, \quad 0 \le d \le (B-1), \quad 1 \le i_s(k) \le L_k,\ -\infty < s < \infty. \tag{9b}$$

The linear combination of the representative vectors is not an inherent property of the Cartesian product of codebooks, but is often applied (e.g., [20]) since the synthesis is linear and high-quality reconstruction of the time domain *waveform* is desired.

We have chosen this specific model mainly because it serves as a *unified framework* for many different coding schemes. This is also the reason for adopting the matrix quantization notation instead of the more popular scalar/vector quantization. For example, the following four particular cases of frequently used scalar/vector quantizers also fall into this general framework.

1) "Vertical" Vector Quantization: This corresponds to $B = J = 1$.

2) Scalar Quantization with Fixed Bit Allocation: This corresponds to $B = 1$, $J = M$ (or $M/2$ for the DFT transform), L_k is the number of quantization levels of the kth sample of the DSTT vector, and $C^{(l,k)}$ is the kth unit vector multiplied by the value of the lth quantization level of the kth sample.

3) "Horizontal" Vector Quantization: This is the generalization of scalar quantization to the case $B > 1$ where $C^{(l,k)}$ is a matrix in which only the kth row is nonzero.

4) Cascaded Vector Quantization [20]: This is a variant of the vertical vector quantization, with $B = 1$, $J > 1$. A coarse vector quantizer of size L_1 is first designed and then the residual vectors $X_{sR} - C^{(i_s)}$ are used as input sequence for a second vector quantizer of size L_2, etc.

In analyzing this class of quantizers, we make the following assumptions.

1) The quantizer is unbiased, i.e.,

$$C^{(l,k)} = E[\mathbf{X}_{s(BR)} | \mathbf{X}_{s(BR)} \in A^{(l,k)}], \quad 1 \le k \le J,\ 1 \le l \le L_k. \tag{10a}$$

This property guarantees an unbiased output signal, and is automatically satisfied by many optimal matrix quantizers (e.g., those designed under a minimum mean-square-error criterion).

2) The expected value of the input sample $x(sBR + d)$ given that $\mathbf{X}_{s(BR)} \in A^{(l,k)}$ depends only on the delay value d, and is known at least for several values of d (as specified in the sequel). Let

$$G^{(l,k)}(d) = E[x(sBR + d) | \mathbf{X}_{s(BR)} \in A^{(l,k)}]. \tag{11a}$$

3) The occurrence probabilities of codewords in each one of the J codebooks, as well as the probabilities of occurrence of a specific pair of codewords, with a specific delay $d(BR)$ between them, are known (estimated[1]), i.e.,

$$P^{(l,k)} = \text{Prob}\{\mathbf{X}_{s(BR)} \in A^{(l,k)}\} \tag{11b}$$

$$F^{(l,k),(i,j)}(d) = \text{Prob}\{\mathbf{X}_{s(BR)} \in A^{(l,k)} \cap \mathbf{X}_{(s+d)(BR)} \in A^{(i,j)}\}. \tag{11c}$$

There are $\hat{L} \triangleq \Sigma_{k=1}^{J} L_k$ different codewords in the union of the J codebooks, and therefore for each value of the

[1]In practice, the frequency of occurrence, or histogram, is measured and is used as an estimate of the probability of occurrence.

delay d, $\hat{L}$ different values of $G^{(\cdot)}(d)$ and $\hat{L}^2$ values of $F^{(\cdot)}(d)$ have to be known.

4) We assume that the input signal $x(n)$ is composed of samples of a wide-sense stationary process with zero mean and known autocorrelation denoted by $\rho(d)$. If the input is a speech signal, $\rho(\cdot)$ represents the long-term autocorrelation sequence of the speech, thus taking advantage of the nonflatness of the speech spectrum in the design process. Furthermore, although in various steps, this sequence appears inside infinite summations, the final result is that only the terms of $\rho(d)$ for $|d| \le L_h + L_f$ are used in the design of the (locally) optimal A/S system, and these terms appear only in the case of FQ. A similar remark applies to the covariance sequence of the additive noise that is used in modeling the FQ effect.

In assumptions 2) and 3) above, we assumed knowledge of values which depend on the codebook used, the analysis window, and type of transform. Therefore, in the sequel, we assume that all these factors have already been determined, and we concentrate on selecting the optimal synthesis filter for a given analysis and coding system (quantizer). Since usually the design of an MQ is based on a typical training sequence, this sequence can also be used to determine $P^{(\cdot)}$, $F^{(\cdot)}$, and $G^{(\cdot)}(d)$ which will afterwards determine the coefficients of the synthesis filter. Again, to simplify the presentation, we will use the sequences $F^{(\cdot)}(d)$ and $G^{(\cdot)}(d)$ inside infinite summations, but in the final result, it is implied that only the values for $|d| \le L_f + L_h$ are actually required.

III. Error Measures

Physically realizable A/S systems introduce signal delay. It is known that the delay of the A/S system depicted in Fig. 1 is an integer multiple of the transform size [1]. We therefore assume that the A/S system has a delay Mr_0, with r_0 being an integer. Since in an ideal system, the reconstructed signal $y(n)$ coincides with the delayed input signal, we define the output error signal as

$$\epsilon(n) \triangleq y(n) - x(n - Nr_0). \tag{12}$$

The assumption that the quantizer is unbiased (7a), (10a) guarantees that the output error signal has zero mean. The WOLA synthesis is a time-varying operation due to the embedded interpolation, and therefore when quantization is applied, $\epsilon(n)$ is *not* a wide-sense stationary process. Thus, in order to measure the error induced by the A/S system with different synthesis filters, we generalize the usual MSE error measure for a class of nonstationary processes which, of course, includes the error process $\epsilon(n)$ above.

Let the autocorrelation sequence of $\epsilon(n)$ be denoted by $\phi(\cdot)$, i.e.,

$$\phi(d, m) \triangleq E[\epsilon(m + d)\, \epsilon(m)] \qquad \forall d, m. \tag{13}$$

In Appendix A, expressions for $\phi(\cdot)$ are given for FQ and MQ. Furthermore, this function has the following two properties (as proven in Appendix A):

$$|\phi(d, m)| \le \text{const} \qquad \forall d, m \tag{14a}$$

$$\phi(d, m + lN) = \phi(d, m) \qquad \forall d, m, l \tag{14b}$$

where $N \triangleq BRM/\gcd(BR, M)$ (where $\gcd(BR, M)$ is the greatest common divisor of the two integer numbers BR and M, and $B = 1$ for the FQ case).

Let $G(f)$ be a *nonnegative* real and symmetric weight function in $L_1[-0.5, 0.5]$ whose Fourier coefficients $g(n) \triangleq \int_{-0.5}^{0.5} G(f)\, e^{j2\pi fn}\, df$ converge absolutely, i.e.,

$$\lim_{l \to \infty} \sum_{n=-l}^{l} |g(n)| < \infty. \tag{15}$$

The first error measure we consider is given by

$$U \triangleq \sum_{d=-\infty}^{\infty} g(d) \left[\frac{1}{N} \sum_{m=0}^{N-1} \phi(d, m) \right] \tag{16}$$

which, due to (14a) and (15), is well defined. Furthermore, it has a spectral domain interpretation as the natural generalization of the MSE criterion for processes with bounded periodic autocorrelation sequence [i.e., satisfying (14)]. The following theorem (whose proof is given in Appendix A) summarizes this interpretation.

Theorem 1: Let

$$u_{l,r} \triangleq \frac{1}{(2r + 1)} \int_{-0.5}^{0.5} \cdot E \left| \sum_{n=l-r}^{l+r} \epsilon(n)\, e^{-j2\pi fn} \right|^2 G(f)\, df \qquad \forall l, r. \tag{17}$$

Then the following hold.

a) For any value of l, $\lim_{r \to \infty} (u_{l,r}) = U$. Therefore, $U \ge 0$.

b) When $\epsilon(n)$ is a wide-sense stationary process with a spectrum $S(f) \in L_2[-0.5, 0.5]$, it follows that

$$U = \int_{-0.5}^{0.5} S(f)\, G(f)\, df. \tag{18}$$

The error measure U defined in (16) is based on the error signal in the time domain. An alternative approach is to define the error in the transform domain (i.e., in the domain in which the quantization is done). Let $\boldsymbol{Y}_{sR}$, $-\infty < s < \infty$ denote the DSTT sequence of vectors generated from the reconstructed signal $y(n)$. In an ideal A/S system, $\boldsymbol{Y}_{SR}$ is a delayed version of the DSTT of the input signal (with delay of Mr_0 samples). Thus, the sequence of error vectors in the transform domain is

$$\boldsymbol{E}_{sR} \triangleq \boldsymbol{Y}_{sR} - \boldsymbol{X}_{(sR - Mr_0)}. \tag{19}$$

Since the DSTT is a linear operation and the error in the time domain $\epsilon(n)$ has zero mean, it follows immediately that $\boldsymbol{E}_{sR}$ also has zero mean. With the DFT transform in mind, it is natural to consider the expected value of the

Euclidean norm of the random error vectors E_{sR}, i.e.,

$$\nu_{sR} \triangleq \mathbf{E}\left[\|E_{sR}\|^2\right] = \sum_{k=0}^{M-1} \mathbf{E}\left[\left|Y_{sR}(k) - X_{sR-Mr_0}(k)\right|^2\right]. \quad (20)$$

It follows from the definition of the DSTT [(1) and (2)] and of $\epsilon(n)$ [in (12)] that ν_{sR} can be represented in terms of the autocorrelation sequence $\phi(\cdot)$ as

$$\nu_{sR} = \sum_{r=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} h(sR - t)\, h(sR - r) \cdot \phi(r - t, t)\, a\left((r)_M, (t)_M\right) \quad (21)$$

where $a(i, j)$ is the (i, j)th element of the $M \times M$ matrix defined as

$$A = T^*T \quad (22)$$

where * denotes "conjugate transpose."

Since the A/S system is time varying, ν_{sR} depends on the time instance sR. However, due to (14b), the sequence ν_{sR} is periodic with a period of N, i.e.,

$$\nu_{(sR+lN)} = \nu_{sR} \qquad \forall s, l \text{ integers}. \quad (23)$$

We now define the second error measure as the time average of ν_{sR} over one period of this sequence, i.e.,

$$V \triangleq \frac{1}{N} \sum_{m=0}^{N-1} \nu_{sR+m}. \quad (24)$$

Substituting (24) in (21) results in

$$V = \sum_{d=-\infty}^{\infty} \left[\sum_{n=-\infty}^{\infty} h(n)\, h(n+d)\right] \cdot \frac{1}{N}\left[\sum_{m=0}^{N-1} \phi(d, m)\, a\left((m+d)_M, (m)_M\right)\right]. \quad (25)$$

Comparing (25) to (16), it is easily verified that when the matrix A is a circulant matrix [i.e., $a(i, j) = a((i + d)_M, (j + d)_M)$ for $d = 0, \cdots, (M - 1)$], V coincides with U, provided that the following weight sequence $g(d)$ is used:

$$g(d) = \sum_{n=-\infty}^{\infty} h(n)\, h(n+d)\, a\left((d)_M, 0\right). \quad (26)$$

It is shown in [14] that the matrix A is circulant for every unitary transform, and then (26) is equivalent to

$$G(f) = \frac{1}{M} \sum_{k=0}^{M-1} \left|H\left(f + \frac{k}{M}\right)\right|^2 \quad (27)$$

where $H(f)$ is the frequency response of the analysis window, and since it is an approximation of an ideal LPF with cutoff frequency $1/2M$, it follows that $G(f) \approx 1$ in (27).

Since the error measure V coincides with U for this wide class of transforms, we continue in what follows with U only.

The error measure U given in (16) is suitable for the design of an optimal synthesis filter, given the analysis window. However, when the design of the optimal analysis window is considered, the error measure U should be modified in order to incorporate the low-pass frequency response specification of the analysis window into the design process. Following the approach in [4], we add the weighted MSE between $H(f)$ and the desired frequency response $D(f)$ to U. Let $W(f) \geq 0$ denote the weight function and $\hat{U}$ the modified error measure; then

$$\hat{U} = U + \int_{-0.5}^{0.5} W(f)\, \left|H(f) - D(f)\right|^2 df. \quad (28)$$

IV. Design of Optimal Synthesis Filters

The optimality criterion is the minimization of U with respect to the unknown synthesis filter coefficients. Combining (16) and the expressions for $\phi(\cdot)$ given in Appendix A, it is easily verified that for both quantization approaches considered here, U is a PSD quadratic form in terms of the synthesis filter vector of coefficients $\boldsymbol{f}$, i.e.,

$$U = C + \frac{1}{R}\left(\boldsymbol{f}'Q\boldsymbol{f} - \boldsymbol{b}'\boldsymbol{f}\right) \quad (29)$$

where the apostrophe denotes transposition. The expressions for the $L_f \times L_f$ matrix Q, the vector $\boldsymbol{b}$ (of dimension L_f), and the constant C are given in Appendix B.

Thus, the optimal synthesis filter is given by the solution of the following set of linear equations:

$$(Q + Q')\boldsymbol{f}_{\text{opt}} = \boldsymbol{b}. \quad (30)$$

If this set of equations is degenerate, each of the infinitely many possible solutions corresponds to the same (minimal) value of U.

We now interpret the general expressions for various particular cases of practical importance, and we analyze both the complexity of the solution and the type of data needed.

Starting with FQ: Here Q can be written as $Q = Q_v + Q_h$ where the matrix Q_v reflects the quantization effect, and the matrix Q_h depends on the analysis window and corresponds to Portnoff's conditions [12].

When no quantization is applied (i.e., $\Psi_{m,n}(d) = 0$), $\hat{x}_{sR}(m) = x_{sR}(m)$. In that case, and if the A/S system is a unity system, $\epsilon(n) = 0$, and therefore $U = 0$. Thus, any unity system is a solution of the (possibly degenerate) set of equations (30), regardless of the weight function $G(f)$ and the input statistics. For $R = M$ and $L_f, L_h > M$ (which is the typical situation in waveform coding applications), *no* unity system exists [13]. However, when $G(f) = 1$, $\rho(d) = \sigma_x^2\delta(d)$, and none of the polyphases [1] of the analysis window is identically zero, (30) possesses a unique solution [14].

For illustration, we show in Fig. 2 the optimal synthesis filter ($L_f = 256$) which was obtained as a solution of (30) for an analysis window which is a truncated sinc sequence (of length $L_h = 256$), approximating an ideal LPF with a cutoff frequency of $1/(2M)$, $M = 16$. In this

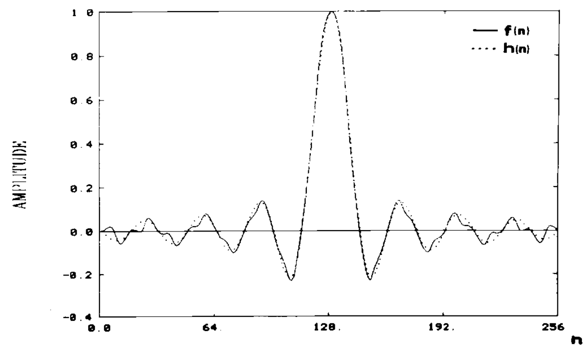


Fig. 2. Optical synthesis filter (solid line) for a given analysis window (dotted line) and an A/S system without quantization ($R = M = 16$, $L_f = L_h = 256$).

example, $G(f) = 1$, $R = M$, $\rho(d) = \delta(d)$ (white noise input). Thus, the uniqueness condition given above is satisfied. The error measure U for this design was numerically evaluated to be -18.17 dB (as elaborated above, a unity system does not exist in this case since here $R = M$ and $L_h > M$).

This A/S system was simulated on a computer, and a speech signal (of telephone bandwidth), sampled at $f_s = 8$ kHz, was passed through it (with no quantization). The output SNR was found to be 16.9 dB and it sounded almost transparent relative to the input speech signal. On the other hand, when the synthesis filter was set identical to the above analysis window, the output SNR dropped to 15.7 dB. Thus, the use of the *optimal* synthesis filter resulted in an improvement of 1.2 dB in output SNR.

If quantization is applied, uniqueness of the solution of (30) is guaranteed, at least when $G(f) = 1$ and the noise process $\boldsymbol{v}_{sR}$ cannot be predicted with zero-mean-square error from $\{\boldsymbol{v}_{(s-d)R}\}_{d=1}^{L_f/R}$ (cf. [14]).

Again, if $Q_v = 0$, (i.e., no quantization is applied), there may be, in general, infinitely many possible solutions of (30), but one can choose among them the appropriate solution assuming a characteristic noise statistic $\hat{\Psi}_{m,n}(dR)$ as follows.

Use $\Psi_{m,n}(\cdot) = \epsilon\hat{\Psi}_{m,n}(\cdot)$ as the noise covariance sequence where $\epsilon > 0$ is a small value governing the overall noise level. If the uniqueness condition presented above is satisfied, then there exists a unique solution of (30) for every $\epsilon > 0$ which is denoted by f_ϵ. It is shown in [14] that $f_0 = \lim_{\epsilon \to 0}(f_\epsilon)$ exists, and that it can be determined efficiently [16] from

$$f_0 = \mathbf{A}'[\mathbf{A}(Q_h + Q_h')\mathbf{A}']^{\div}\mathbf{A}b \tag{31}$$

where $\mathbf{A}\mathbf{A}' = \hat{Q}_v + \hat{Q}_v'$ [for $\hat{Q}_v$ that is generated by $\hat{\Psi}_{m,n}(\cdot)$] and $\div$ denotes the Moore–Penrose (M–P) inverse [15].

In order to illustrate the effect of the quantization noise, we present the following two examples for which a *closed-form* solution of (30) can be obtained.

Example 1: $G(f) = 1$, $L_h = L_f = M$, $r_0 = 1$, $\rho(d) = \sigma_x^2\delta(d)$, $\Psi_{m,n}(dR) = \delta(d)\,\hat{\Psi}((m-n)_M)$. In this case, it is shown in Appendix B that the optimal synthesis filter is

$$f_{\text{opt}}(t) = \frac{h(M-t)}{\sum_{r=-\infty}^{\infty} h(M-t-rR)^2 + \hat{\Psi}(0)/\sigma_x^2}, \quad 0 \le t \le M-1 \tag{32}$$

which extends the results in [5], [6] to include the effect of quantization noise.

Example 2: $G(f) = 1$, $R = M$. In this case, the set of L_f linear equations given in (30) is decomposable into M sets, of about the same number of equations in each. Let $f_\tau(x)$ denote the τth polyphase [1] of the synthesis filter (i.e., $f_\tau(x) \triangleq f(\tau + Mx)$, $0 \le \tau \le M-1$); $h_\tau(x)$ denotes the properly delayed and inverted in time, τth polyphase of the analysis window (i.e., $h_\tau(x) = h(Mr_0 - (\tau + Mx))$, $0 \le \tau \le M-1$), and $\tilde{\rho}(d)$ denotes the decimated by M autocorrelation sequence of the input signal. Then, the optimal synthesis filter is the solution of

$$\begin{aligned}\sum_{d=-\infty}^{\infty} & h_\tau(x-d)\,\tilde{\rho}(d) \\ &= \sum_{y=-\infty}^{\infty} f_\tau(x-y)[R_\tau(y) + \Psi_{\tau,\tau}(yM)], \\ & 0 \le \tau \le M-1 \\ & 0 \le x \le \left\lfloor \frac{L_f - 1 - \tau}{M} \right\rfloor\end{aligned} \tag{33}$$

where

$$R_\tau(y) \triangleq \sum_{r=-\infty}^{\infty} h_\tau(r+y) \sum_{l=-\infty}^{\infty} h_\tau(r-l)\,\tilde{\rho}(l), \quad -\infty < y < \infty. \tag{34}$$

The derivation of (33), (34) is given in Appendix B. For $L_f = \infty$, and if noncausality of the synthesis filter is allowed (i.e., (33) holds for $-\infty < x < \infty$), these equations have a frequency domain interpretation which resembles the classical Wiener filter. Let $F_\tau(f)$, $H_\tau(f)$, $\tilde{R}(f)$, $\Psi_{\tau,\tau}(f)$ be the frequency responses of $f_\tau(x)$, $h_\tau(x)$, $\tilde{\rho}(d)$, and $\Psi_{\tau,\tau}(dM)$, respectively; then (33), (34) imply that

$$F_\tau(f) = \frac{H_\tau(f)\,\tilde{R}(f)}{\Psi_{\tau,\tau}(f) + \tilde{R}(f)\,|H_\tau(f)|^2}, \quad -0.5 < f \le 0.5. \tag{35}$$

The complexity of the design according to (33), (34) is quite small, as only $(L_f + L_h)/M$ values of $\rho(\cdot)$ and L_f/M values of $\Psi_{m,m}(\cdot)$ are used and only M Toeplitz systems of about L_f/M linear equations each have to be solved. A more general complexity analysis is given in [14].

For illustration, Fig. 3 depicts the optimal synthesis filter (solid line) obtained for an A/S system with a

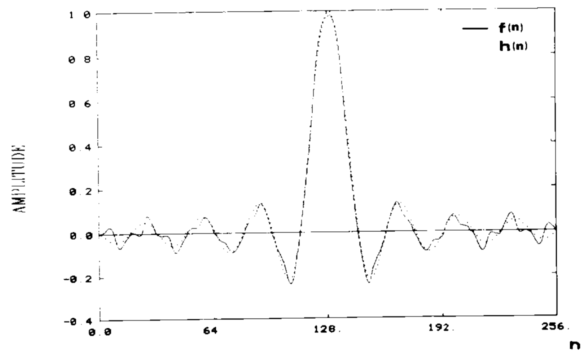


Fig. 3. Optimal synthesis filter (solid line) for a given analysis window (dotted line) and an A/S system with a DFT and fine quantization (transmission rate = 32 kbits/s).

DFT and "fine quantization." The design is based on (33), (34) with L_h, $L_f = 256$; $R = M = 16$, and it is based on statistical data estimated from 10 s of speech ($f_s = 8$ kHz). The following simple quantization scheme is used in this example.

The quantizer in each band is uniform (its step size being optimized for the Gamma distribution). The real and imaginary components of each transformed sample are separately encoded (only half of the complex transform components are encoded and transmitted since the input signal is real). The number of bits assigned to each band was fixed and was set according to the long-term variance estimate of the transformed samples (to minimize the MSE, but limited to a maximum of 6 bits/component). The overall transmission bit rate was set in this example to 32 kbits/s (i.e., 64 bits per each transformed vector).

The main computational load in the design is the estimation of the statistical data needed in (33), (34), namely, $\tilde{\rho}(d)$ and $\Psi_{\tau,\tau}(dM)$. For 10 s of speech (80 000 samples), it required about $5 \cdot 10^6$ multiply-add operations (evenly split between the two estimated functions). Adding the operations needed to complete (34) and to solve the $M = 16$ Toeplitz systems of $L_f/M = 16$ equations, less than $5.2 \cdot 10^6$ multiply-add operations were needed to complete the design of the optimal synthesis window in this example. An output SNR of 15 dB was measured with the above system, compared to 14.1 dB when a nonoptimal sinc synthesis filter (of length $L_f = 256$) is used.

We turn now to MQ.

In this case, due to the inherent nonlinearity of the quantization process, the matrix Q cannot be decomposed into $Q_h + Q_r$ as with FQ. Moreover, the implicit *dependence* of the synthesis filter on the given analysis window coefficients is only via the statistics $F^{(\cdot)}(d)$, $G^{(\cdot)}(d)$, and $P^{(\cdot)}$. Furthermore, since there are only L possible different MDSTT vectors, it is clear that, in general, no unity system exists (i.e., $U > 0$ even for f_{opt} obtained via (30), and no matter what are the values of R, M, L_h, L_f). A sufficient condition for the uniqueness of f_{opt} obtained from (30) is (as for FQ) $G(f) = 1$, and the process

$$f(m) = \frac{\sum_{d=0}^{B-1}\sum_{l=1}^{\hat{L}} P^{(l)} c_d^{(l)}(m)\, G_d^{(l)}\left(m + M(d - r_0)\right)}{\sum_{d=0}^{B-1}\sum_{k=1}^{\hat{L}}\sum_{l=1}^{\hat{L}} F^{(k),(l)}(0)\, c_d^{(k)}(m)\, c_d^{(l)}(m)},$$

$$0 \le m \le M - 1 \tag{36}$$

where for ease of presentation, we reorder the $\hat{L}$ codewords and thus replace the pairs of indexes (l, k) in (11) by a single index.

This result is derived in Appendix B, as well as the simplified expressions of Q and $\boldsymbol{b}$ for the important case of $G(f) = 1$, $R = M$, as well as a discussion on the complexity of the design.

For illustration, we show in Fig. 4 the optimal synthesis filter (solid line) obtained for an A/S system with a DFT and vertical vector quantization [i.e., $B = J = 1$—see in Section II] of size $M = 8$ (here also $R = M$ and $G(f) = 1$). The synthesis filter is obtained by solving (30), using (B7a), (B7b) which are simplified further here to have the form

$$b(t) = 2\sum_{l=1}^{\hat{L}} P^{(l)} c^{(l)}\left((t)_M\right) G^{(l)}(t) \tag{37a}$$

$$Q(t, s) = \sum_{l=1}^{\hat{L}}\sum_{k=1}^{\hat{L}} c^{(l)}\left((t)_M\right) c^{(k)}\left((t)_M\right) \cdot F^{(l),(k)}\left(\frac{s-t}{M}\right) \delta\left((s-t)_M = 0\right)$$

$$0 \le t, s \le L_f - 1. \tag{37b}$$

Furthermore, with $Q(t, s)$ above, the matrix $[Q + Q^T]$ [in (30)] has zeros everywhere except on diagonals separated M elements apart (including the main diagonal). Thus, the synthesis filter can be solved here in terms of its M polyphase filters, with each polyphase being solved from a set of L_f/M equations.

Because of speech nonstationarity a *gain-adaptive* vector quantization scheme [21] is used with forward adaptation. With 4 bits for the gain (on a log scale) computed for every 16 consecutive vectors (corresponding to 0.25 bits/vector), and a dictionary of 256 vectors (i.e., 8 bits/vector), the overall transmission rate is 8.25 kbits/s. Since here the effective value of $\hat{L}$ is $2^4 \cdot 2^8 = 2^{12}$, storing $F^{(l),(k)}(\cdot)$ requires a large amount of storage. We have chosen therefore to directly generate $b(t)$ and $Q(t, s)$ by updating its values as each input vector from the training data is entered. The updating equations turn out to be

$$\bar{b}(t) \leftarrow \bar{b}(t) + 2\hat{x}_{sR}\left((t)_M\right) x(t + sR),$$

$$0 \le t \le L_f - 1 \tag{38a}$$

$$\bar{Q}(t, r) \leftarrow \bar{Q}(t, r) + \hat{x}_{sR}\left((t)_M\right) \hat{x}_{[s+(r-t)/M]R}\left((t)_M\right),$$

$$0 \le t, r \le L_f - 1; \quad (r - t)_M = 0 \tag{38b}$$

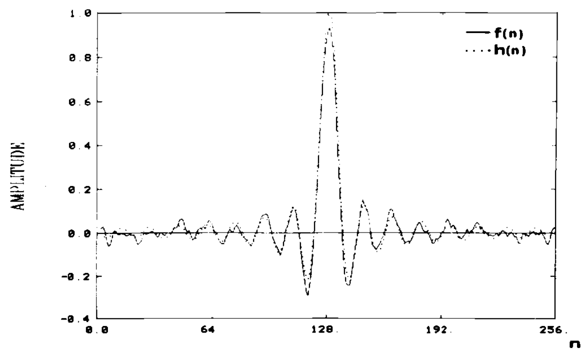


Fig. 4. Optimal synthesis filter (solid line) for a given analysis window (dotted line) and an A/S system with gain-adaptive "vertical" vector quantization (transmission rate = 8.25 kbits/s).

where $x(t + sR)$ is a sample at time $(t + sR)$ of the input signal, $\hat{x}_{sR}(t)$ is the tth element of the quantized (reproduction) vector $\hat{\boldsymbol{x}}_{sR}$, and $\bar{b}(t) = \bar{N}b(t)$, $\bar{Q}(t, r) = \bar{N}Q(t, r)$ where $\bar{N}$ is the number of vectors used for training.

For $\bar{N} = 10^4$ training vectors (input data of $8 \cdot 10^4$ samples) and $L_f = 256$, about $3\bar{N}L_f = 7.68 \cdot 10^6$ multiply-add operations are needed for computing $b(t)$ and $Q(t, r)$ in this case. The solution of $M = 8$ Toeplitz systems of $L_f/M = 32$ equations each is of much less complexity ($O(10^5)$). The amount of computations needed for the analysis, vector quantization process (search in dictionary), and WOLA synthesis are not included here as they are part of the usual coding system.

The above system (with the optimal synthesis filter and 8.25 kbit/s transmission rate) resulted in an output SNR of 12 dB (for 10 s of input speech) and in 11.5 dB when the synthesis filter was set to be identical to the analysis window (i.e., truncated sinc sequence of length $L_f = L_h = 256$). Although only 0.5 dB improvement was obtained in this particular experiment,[2] a clear subjective improvement in reproduced speech quality could be discerned, apparently because of the different (smoother) nature of the error.

It should be noted that the above scheme ("vertical" VQ) is particularly attractive if the squared error distance measure is used to construct the dictionary and to search it for the best reproduction vector. This is because there is actually no need to perform the transform at the transmitter and the inverse transform at the receiver if the transform is unitary (which the DFT is). The encoding can thus be done here on the time-domain vectors $\boldsymbol{x}_{sR}$ directly.

V. Iterative Design of Optimal A/S Systems with Fine Quantization

If fine quantization (FQ) is used, the error measure $\hat{U}$ is given as an explicit function of the coefficients of the analysis window. Combining (16) and (28) and the expression for $\phi(\cdot)$ given in Appendix A, it is easily verified that $\hat{U}$ is a PSD quadratic form in terms of the analysis window vector of coefficients $\boldsymbol{h}$. Thus, *for a given synthesis* filter, the optimal analysis window which minimizes $\hat{U}$ is the solution of a linear set of equations, as follows:

$$(\tilde{Q}_f + \tilde{Q}_f' + \tilde{Q}_D + \tilde{Q}_D')\boldsymbol{h}_{\text{opt}} = \tilde{\boldsymbol{b}}_D + \tilde{\boldsymbol{b}}_f. \quad (39)$$

The expressions for the $L_h \times L_h$ matrices $\tilde{Q}_f$, $\tilde{Q}_D$ and the vectors $\tilde{\boldsymbol{b}}_D$, $\tilde{\boldsymbol{b}}_f$ of dimension L_h are given in Appendix C. It is also shown there that for a weight function $W(f)$ which is positive on a set of nonzero measure, there exists a unique solution of (37). The matrix $\tilde{Q}_D$ and the vector $\tilde{\boldsymbol{b}}_D$ reflect the frequency response specification on the analysis window via (28), whereas $\tilde{Q}_f$ and $\tilde{\boldsymbol{b}}_f$, which depend on the synthesis filter, reflect the desired unity system specification. The statistics of the quantization noise *do not affect* the optimal analysis window. Furthermore, regardless of the quantization, the optimal analysis window *does not correspond in general to a unity system*, but is rather a compromise between the frequency response specification and the unity system specification. Unlike the set of linear equations in (30), the equations in (37) are usually irreducible even for $M = R$ and $G(f) = 1$.

The design of an optimal A/S system with FQ is done by the following iterative algorithm, similar to [4].

1) Initialize $r = 0$, and let $\boldsymbol{h}^{(0)}$, $\boldsymbol{f}^{(0)}$ denote the given initial analysis and synthesis filters.

2) Let $\boldsymbol{h}^{(r+1)} \in A_1(\boldsymbol{f}^{(r)})$ be any solution of (37) for the given synthesis filter $\boldsymbol{f}^{(r)}$, with the exception that if $\boldsymbol{h}^{(r)}$ is a solution, choose this solution.

3) Let $\boldsymbol{f}^{(r+1)} \in A_2(\boldsymbol{h}^{(r+1)})$ be any solution of (30) for the given analysis window $\boldsymbol{h}^{(r+1)}$, with the exception that if $\boldsymbol{f}^{(r)}$ is a solution, choose this solution.

4) If $\boldsymbol{f}^{(r)} = \boldsymbol{f}^{(r+1)} \wedge \boldsymbol{h}^{(r)} = \boldsymbol{h}^{(r+1)}$, stop[3]; otherwise, $r \leftarrow (r + 1)$ and return to 2).

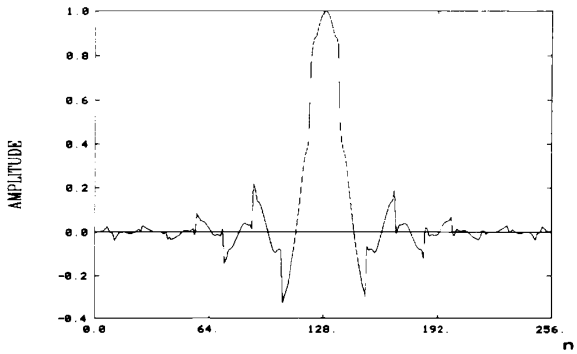
Here $A_1(\cdot)$ and $A_2(\cdot)$ denote the sets of solutions of (37) and (30), respectively.

This iterative algorithm consists of alternately solving two sets of linear equations. For the special case of $R = M$, $G(f) = 1$, $\Psi_{m,n}(d) = 0$, and $\rho(d) = \delta(d)$, $\hat{U}$ is similar to the error measure in [4] which is the deterministic MSE between the A/S system unit sample response and the desired ideal response. The only difference is that we incorporated the linear constraint on the gain of the system, used in [4] into the error measure, thus avoiding the need of using Lagrange multipliers, and hence simplifying the design algorithm.

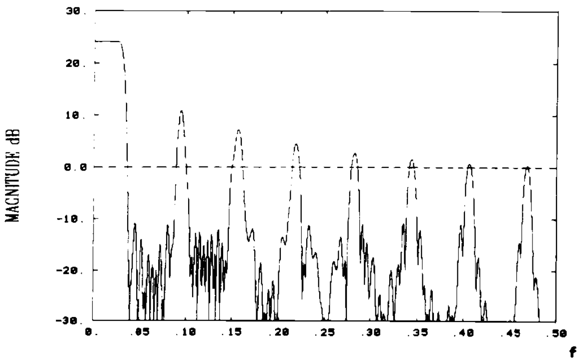
For illustration, Fig. 5(a) shows the result of an iterative design of an analysis–synthesis window pair [for $R = M = 16$, $L_f = L_h = 256$, $G(f) = 1$, $\rho(d) = \delta(d)$] *without quantization* and without any constraint on the

[2]An additional improvement of 0.4 dB was obtained when the sinc analysis window was replaced by a window such as the one in Fig. 5 (but for $M = 8$) which was designed by applying the iterative technique described in the next section.

[3]In practice, since the arithmetic has final precision, the number of iterations is not allowed to exceed a given limit. Furthermore, the iterations are stopped whenever $U(\boldsymbol{f}^{(r)}, \boldsymbol{h}^{(r)}) - U(\boldsymbol{f}^{(r+1)}, \boldsymbol{h}^{(r+1)}) \leq \epsilon$. However, to obtain convergence proofs, we avoid this issue in the discussion above.



(a)



(b)

Fig. 5. (a) Analysis/synthesis window pair obtained by iterative design (identical windows) for an A/S system without quantization and with $R = M = 16$, $L_f = L_h = 256$. (b) Frequency response of the window in Fig. 5(a).

analysis window frequency response (i.e., $W(f) = 0$, $\hat{U} = U$). In this case, the analysis and synthesis windows are *identical*, and they result in a system which is very close to a unity system ($U < -60$ dB, SNR > 60 dB). The frequency response of each of these windows is shown in Fig. 5(b). It is of interest to note that convergence to the same pair of windows shown in Fig. 5(a) was obtained from several initial analysis windows (of length $L_h = 256$) having a low-pass response with the proper cutoff frequency. Convergence was obtained within 50–100 iterations.

Using the above pair of windows in an A/S system with fine quantization (and DFT) resulted in an output SNR of 18.6 dB at the bit rate of 32 kbits/s. This is a marked improvement over the results obtained with the pair of windows depicted in Fig. 3. Further improvement could be expected by adding a constraint on the frequency response of the analysis window by using a nonzero value for $W(f)$ in (28), and by taking into account the quantization noise statistics, as well as using gain adaptation (as was done with VQ above).

The following theorem which is proved in Appendix C summarizes the convergence properties of the iterative algorithm, assuming both $\tilde{Q}_D$ and Q_v are PD matrices (i.e., both (37) and (30) always possess unique solution). The reader may consider [14] for a more complete version of this theorem.

Theorem 2:

a) $\hat{U}$ is monotonically decreasing from iteration to iteration, unless the algorithm stops at a fixed point in $\Gamma \triangleq \{(\boldsymbol{f}, \boldsymbol{h}); \boldsymbol{h} = A_1(\boldsymbol{f}) \cap \boldsymbol{f} = A_2(\boldsymbol{h})\}$.

b) Γ is the set of stationary points of $\hat{U}$.

c) Every sequence generated by the algorithm has at least one limit point, and all limit points are in Γ.

d) $\hat{U}$ possesses a global minimum which is in Γ.

Assume that $(\boldsymbol{f}^*, \boldsymbol{h}^*) \in \Gamma$; then the iterative algorithm converges linearly to $(\boldsymbol{f}^*, \boldsymbol{h}^*)$, and the rate of convergence is determined by the second-order derivatives of $\hat{U}$ in $(\boldsymbol{f}^*, \boldsymbol{h}^*)$ (cf. [14]).

VI. Conclusions

In this paper, a statistical model of analysis/synthesis systems with quantization is presented. It is applied to the design of systems under two different quantization approaches: fine quantization and matrix quantization (which serves as a unified framework having scalar and vector quantization as special cases), and any linear regular transform. With the presented statistical model, two different error measures were defined: the first one in the time domain, and the second one in the transform domain. These error measures are shown to be equivalent for a wide class of transforms (including the DFT). The optimal WOLA synthesis filter is obtained by solving a set of linear equations. Conditions for uniqueness of the solution when quantization is applied are given, and it is shown that if there is no quantization, then an optimal synthesis filter corresponds to a unity system (if it exists). Thus, for $R < M$, the result in [5], [6] is just one of many possible optimal synthesis filters. Furthermore, this result is extended to the case of white input signal and white (nonzero) quantization noise. It is also shown that for quantization noise approaching zero, there exists a limiting optimal synthesis filter which corresponds to a unity system, but depends on the specific input signal and noise statistics.

For the important case of critical sampling ($R = M$), the polyphases of the optimal synthesis filter are interpreted in the frequency domain as Wiener filters. For this case, the new approach partially extends the results of [4] by incorporating the quantization effect into the design process, and thus leads to optimal synthesis filters which *depend on the quantization characteristics*.

For the case of fine quantization, an iterative algorithm (similar to [4]) is presented for the design of optimal A/S system, and its convergence properties are analyzed.

Preliminary computer simulations with speech coding have shown both objective (SNR) and subjective (by informal listening) improvement as compared to A/S systems with nonoptimal windows, particularly at low bit rates.

Appendix A

A. Explicit Equations for $\phi(d, m)$ and Its Properties

Following (12), (13), and (4), it follows that

$$\phi(d, m) = \rho(d) - \sum_{s=-\infty}^{\infty} f(m - sR)\, E\big[\hat{x}_{sR}((m)_M) \cdot x(m + d - Mr_0)\big] - \sum_{s=-\infty}^{\infty} f(m + d - sR)\, E\big[\hat{x}_{sR}((m + d)_M) \cdot x(m - Mr_0)\big] + \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} f(m - tR) f(m + d - sR) \cdot E\big[\hat{x}_{sR}((m + d)_M)\, \hat{x}_{tR}((m)_M)\big]. \quad \text{(A1)}$$

Following (1), (6), and (7), we obtain for the *FQ case*

$$E\big[\hat{x}_{sR}((m)_M)\, x(m + d - Mr_0)\big] = \sum_{r=-\infty}^{\infty} h(sR - m - rM)\, \rho(M(r + r_0) - d) \quad \text{(A2a)}$$

$$E\big[\hat{x}_{sR}((m + d)_M)\, x(m - Mr_0)\big] = \sum_{r=-\infty}^{\infty} h(sR - m - d - rM)\, \rho(M(r + r_0) + d) \quad \text{(A2b)}$$

$$E\big[\hat{x}_{sR}((m + d)_M)\, \hat{x}_{tR}((m)_M)\big] = \Psi_{(m+d)_M,(m)_M}((t - s)R) + \sum_{r=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \cdot h(tR - m - rM)\, h(sR - m - d - nM) \cdot \rho(d + M(n - r)). \quad \text{(A2c)}$$

Following (9)–(11), we obtain for the *MQ case*

$$E\big[\hat{x}_{sR}((m)_M)\, x(m - d - Mr_0)\big] = \sum_{l=1}^{\hat{L}} P^{(l)} c^{(l)}_{(s)_B}((m)_M) \cdot G^{(l)}\left(m + d - Mr_0 - \left\lfloor \frac{s}{B} \right\rfloor BR\right) \quad \text{(A3a)}$$

$$E\big[\hat{x}_{sR}((m + d)_M)\, x(m - Mr_0)\big] = \sum_{l=1}^{\hat{L}} P^{(l)} c^{(l)}_{(s)_B}((m + d)_M) \cdot G^{(l)}\left(m - Mr_0 - \left\lfloor \frac{s}{B} \right\rfloor BR\right) \quad \text{(A3b)}$$

$$E\big[\hat{x}_{sR}((m + d)_M)\, \hat{x}_{tR}((m)_M)\big] = \sum_{l=1}^{\hat{L}} \sum_{k=1}^{\hat{L}} F^{(l),(k)}\left(\left\lfloor \frac{t}{B} \right\rfloor - \left\lfloor \frac{s}{B} \right\rfloor\right) \cdot c^{(l)}_{(s)_B}((m + d)_M)\, c^{(k)}_{(t)_B}((m)_M). \quad \text{(A4)}$$

All the infinite summations in (A1) and (A2) are actually finite since both $f(\cdot)$ and $h(\cdot)$ are of finite length. The periodicity of $\phi(d, m)$ in the second variable with a period of $N \triangleq BRM/\gcd(BR, M)$ can be verified from (A1)–(A4) using elementary (although tedious) algebra. For the boundness of $|\phi(d, m)|$, it is enough to check that this sequence is bounded with respect to d for $0 \le m \le N - 1$. This property follows from the fact that $f(\cdot)$ and $h(\cdot)$ are of finite values and length, and from the assumption that the covariance sequences $\rho(\cdot)$ and $\Psi(\cdot)$ are bounded, as well as the conditional expectations sequences $G^{(l)}(\cdot)$. Note that the terms which appear in (A4) are always bounded due to the boundness of $0 \le F^{(l),(k)}(\cdot) \le 1$. The boundness assumptions on the input signal are quite natural and not very restrictive.

B. Proof of Theorem 1

1) Since the summation over n in (17) is finite, all the expectations involved are finite [due to (14a)], and the Fourier coefficients of $G(f)$ are well defined, it follows that

$$u_{l,r} = \sum_{n=l-r}^{l+r} \sum_{m=l-r}^{l+r} g(n - m) \cdot \phi(n - m, m) \frac{1}{(2r + 1)}. \quad \text{(A5)}$$

Using the periodicity of $\phi(\cdot)$ with respect to the second variable [see (14b)], we rewrite (A5) as follows:

$$u_{l,r} = \sum_{d=-2r}^{2r} g(d) \sum_{m=0}^{N-1} \phi(d, m)\, w_{l,r}(d, m) \quad \text{(A6)}$$

with

$$w_{l,r}(d, m) \triangleq \Big|\{x;\ l - r \le x \le l + r \wedge l - r \le x + d \le l + r \wedge x \equiv m (\mathrm{mod}\ N)\}\Big| / (2r + 1).$$

The sequence $w_{l,r}(d, m)$ is a window sequence having the following two properties:

$$w_{l,r}(d, m) = 0 \qquad |d| \ge (2r + 1) \quad \text{(A7a)}$$

$$\left|\frac{1}{N}\left(1 - \frac{|d|}{2r + 1}\right) - w_{l,r}(d, m)\right| \le \frac{1}{(2r + 1)} \qquad |d| \le 2r. \quad \text{(A7b)}$$

Define

$$\nu_r \triangleq \sum_{d=-2r}^{2r} g(d)\left|\frac{1}{N}\sum_{m=0}^{N-1}\phi(d,m)\right|\left(1-\frac{|d|}{2r+1}\right)$$

$$u_{2r} \triangleq \sum_{d=-2r}^{2r} g(d)\left|\frac{1}{N}\sum_{m=0}^{N-1}\phi(d,m)\right|. \tag{A9}$$

Then, from (A7a), (A7b), (A8), (A6), (14a), and the triangle inequality for the l_1 norm,

$$|\nu_r - u_{l,r}| \le \sum_{d=-2r}^{2r} |g(d)| C \frac{1}{(2r+1)}. \tag{A10}$$

Due to the Riemann–Lebesgue lemma [18], $\lim_{d\to\infty} |g(d)| = 0$ for every $G(f) \in L_1[-0.5, 0.5]$. Therefore, from (A10), it follows that $\lim_{r\to\infty} |\nu_r - u_{l,r}| = 0$. Thus, it is sufficient to show that $\lim_{r\to\infty} \nu_r = U$, and since $\nu_r = (1/2r+1)\sum_{k=0}^{2r} u_k$, $\lim_{r\to\infty} u_r = U$ implies that $\lim_{r\to\infty}\nu_r = U$. Let

$$\tilde{\phi}(d) = \frac{1}{N}\sum_{m=0}^{N-1}\phi(d,m). \tag{A11}$$

Since $\sum_{d=-\infty}^{\infty} g(d)$ converges absolutely, for $\tilde{\phi}(d)$ which is bounded [by (14a)], $\sum_{d=-\infty}^{\infty} g(d)\tilde{\phi}(d)$ also converges (absolutely), and therefore $\lim_{r\to\infty} u_r = U < \infty$. This completes part 1) of the proof.

2) For $\epsilon(n)$ which is a wide-sense stationary process, the period N of $\phi(\cdot)$ is one. Therefore, from (16), $U = \sum_{d=-\infty}^{\infty} g(d)\phi(d)$. Since $g(d)$ converges absolutely, it follows that $G(f) \in L_\infty[-0.5, 0.5] \subseteq L_2[-0.5, 0.5]$. If $\epsilon(n)$ has a spectrum $S(f)$ which is in $L_2[-0.5, 0.5]$, we can apply Parseval's theorem to obtain (18) (due to the realness of $G(f)$, the complex conjugate operation is omitted).

Appendix B

A. The Expressions of Q, b, C in (29)

The quadratic form (29) is obtained by substituting (A1)–(A4) in (16) and rearranging the resulting expression as a quadratic form in terms of the coefficients of the synthesis filter. We present here only the final results, i.e., the expressions for computing the matrix Q, the vector $\boldsymbol{b}$, and the scalar C.

$$C = \sum_{d=-\infty}^{\infty} \rho(d)\, g(d). \tag{B1}$$

For the *FQ case*, the tth element of the vector $\boldsymbol{b}$ ($0 \le t \le L_f - 1$) is

$$b(t) = \sum_{r=-\infty}^{\infty} h(Mr-t) \sum_{d=-\infty}^{\infty} g(d) \cdot \big(\rho(d - M(r-r_0)) + \rho(-d - M(r-r_0))\big) \tag{B2a}$$

and the elements of the matrix Q are given by

$$Q(t,s) = \sum_{r=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} h(Mr-s)\,h(Mn-t) \cdot \sum_{d=-\infty}^{\infty} \rho\big(M(n-r)+(s-t)+Rd\big) \cdot g(s-t+Rd) + \sum_{d=-\infty}^{\infty} g(s-t+Rd) \cdot \frac{R}{N}\sum_{\substack{m=0\\ m\equiv(t)_R}}^{(N-1)} \Psi_{(m+(s-t)+Rd)_M,(m)_M}(-Rd),$$

$$0 \le t, s \le L_f - 1. \tag{B2b}$$

Likewise, for the *MQ case*, we obtain

$$b(t) = \sum_{j=0}^{B-1}\sum_{l=1}^{L} P^{(l)} \frac{R}{N} \sum_{\substack{m=0\\ (m-t)_R=0\\ (m-t)_{BR}=j}}^{N-1} c_j^{(l)}((m)_M) \cdot \sum_{d=-\infty}^{\infty} g(d)\big(G^{(l)}(d - Mr_0 + t + jR) + G^{(l)}(-d - Mr_0 + t + jR)\big) \tag{B3a}$$

$$Q(t,s) = \sum_{j=0}^{B-1}\sum_{l=1}^{L}\sum_{k=1}^{L} \frac{R}{N} \sum_{\substack{m=0\\ (m-t)_R=0\\ (m-t)_{BR}=j}}^{N-1} \cdot \sum_{d=-\infty}^{\infty} c^{(l)}_{(j+d)_B}\big((m+(s-t)+Rd)_M\big)\, c_j^{(k)}((m)_M) \cdot F^{(l),(k)}\left(-\left\lfloor\frac{j+d}{B}\right\rfloor\right) g(s-t+Rd). \tag{B3b}$$

The indexes (t, s) in (B2), (B3) are limited to a finite interval of length L_f, for which the synthesis filter's coefficients are nonzero. The summations over the (r, n) indexes in (B2) are therefore finite, but the summations over d in (B2), (B3) may be infinite if $g(\cdot)$ is infinite. However, these summations can be alternatively evaluated in the frequency domain using Parseval's theorem. The matrix Q in (B2b) is interpreted throughout as the sum of two matrices $Q_h + Q_\nu$ where $Q_h(t, s)$ is the first term on the right-hand side of (B2b) and $Q_\nu(t, s)$ is the second term.

B. Derivations for Example 1

We substitute in (B2a) $g(d) = \delta(d)$ (since $G(f) = 1$) and $\rho(d) = \delta(d)\sigma_x^2$, and the summations over d and r collapse into a single element $d = 0$, $r = r_0 = 1$, thus leading to $b(t) = 2h(M-t)\sigma_x^2$. Substituting these values and $\Psi_{m,n}(dR) = \delta(d)\,\hat{\Psi}((m-n)_M)$ into (B2b), we notice that $Q(t, s)$ equals zero unless $(t-s)_R = 0$.

For $s = t + Rr$, it follows that

$$Q(t, t + rR) = \sigma_x^2 \sum_{n=-\infty}^{\infty} h(Mn - t - rR) \cdot h(Mn - t) + \hat{\Psi}(0)\,\delta(r). \quad \text{(B4)}$$

Now, since $L_h = L_f = M$, the summation over n in (B4) can be replaced by a single element $n = 1$, and therefore in this example, (30) becomes

$$h(M - t) \sum_{r=-\infty}^{\infty} h(M - t - rR) f(t + rR) + \left(\frac{\hat{\Psi}(0)}{\sigma_x^2}\right) f(t) = h(M - t), \quad 0 \le t \le (M - 1). \quad \text{(B5)}$$

It is easily verified that the unique solution of (B5) for $\hat{\Psi}(0) > 0$ is given by (32), and that for $\hat{\Psi}(0) = 0$, it coincides with the limiting synthesis filter f_0 for white noise.

C. *Derivations for Example 2*

When we substitute into (B2) $R = M = N$ and $g(d) = \delta(d)$, we obtain

$$b(t) = 2 \sum_{x=-\infty}^{\infty} h(Mr_0 - (t + Mx))\,\rho(Mx) \quad \text{(B6a)}$$

$$Q(t, s) = \left\{ \sum_{r=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} h(Mn - t)\, h(Mr - s) \cdot \rho(M(n - r)) + \Psi_{(t)_M,(t)_M}(s - t) \right\} \cdot \delta((s - t)_M = 0). \quad \text{(B6b)}$$

It follows from (B6b) that the set of L_f linear equations given in (30) is decomposable into M sets, each of them of L_f/M equations. Using the notations of $f_\tau(x)$, $h_\tau(x)$ and $\tilde{\rho}(d) = \rho(Md)$ and the even symmetry of the sequences $\tilde{\rho}(d)$ and $\Psi_{\tau,\tau}(d\,M)$, we can obtain (33), (34) from (30) and (B6).

D. *Simplified Equations for the MQ Case when $M = R$, $G(f) = 1$*

When $M = R$ and $G(f) = 1$, (B3) can be significantly simplified and yields

$$b(t) = \frac{2}{B} \sum_{j=0}^{B-1} \sum_{l=1}^{\hat{L}} P^{(l)} c_j^{(l)}((t)_M) \cdot G^{(l)}(t + (j - r_0)M) \quad \text{(B7a)}$$

$$Q(t, s) = \frac{1}{B} \sum_{j=0}^{B-1} \sum_{l=-1}^{\hat{L}} \sum_{k=-1}^{\hat{L}} c^{(l)}_{(j+(t-s)/M)_B}((t)_M) \cdot c_j^{(k)}((t)_M)\, F^{(l),(k)}\left(-\left\lfloor \frac{j + (t - s)/M}{M} \right\rfloor\right) \cdot \delta((s - t)_M = 0). \quad \text{(B7b)}$$

For $L_f = M$, the matrix Q is diagonal, and (36) follows from substituting (B7) in (30). For $L_f > M$, there is no closed form solution, but yet the set of L_f linear equations is decomposable into M sets of about (L_f/M) equations each, and each of these sets is represented by a Toeplitz matrix. Therefore, the complexity of solving (30) is $\mathbf{O}((L_f/M)^2 M)$. The design process involves $\hat{L}^2(L_f/MB)$ different values of $F^{(l),(k)}(d)$, and $\hat{L}(L_f + B)$ different values of $G^{(l)}(d)$. The overall complexity of evaluating Q and b is $\mathbf{O}(B\hat{L}^2 L_f + B\hat{L}L_f)$. The complexity analysis of the more general cases in which either $R \ne M$ or $G(f) \ne 1$ is omitted here for simplicity.

Appendix C

A. *Expressions for $\tilde{Q}_D$, $\tilde{Q}_f$, $\tilde{b}_D$, $\tilde{b}_f$ in (37)*

The expressions for the matrices $\tilde{Q}_D$ and $\tilde{Q}_f$, of dimensions $L_h \times L_h$, and for the vectors $\tilde{b}_D$, $\tilde{b}_f$, of dimension L_h, are obtained by substituting (A1)–(A2) into (16), combining it with (28), and rearranging the results as a quadratic form in terms of the analysis window coefficients. For simplification, we omit here the details of this easy (although lengthy) derivation, and only state the final results, which are

$$\tilde{b}_D(t) = 2\,\mathrm{Re}\left\{ \int_{-0.5}^{0.5} W(f)\, D(f)\, e^{j2\pi ft}\, df \right\} \quad \text{(C1a)}$$

$$\tilde{b}_f(t) = \frac{1}{R} \sum_{r=-\infty}^{\infty} f(Mr - t) \sum_{d=-\infty}^{\infty} g(d) \cdot \left(\rho(d - M(r - r_0)) + \rho(-d - M(r - r_0))\right) \quad \text{(C1b)}$$

$$\tilde{Q}_D(t, s) = \left\{ \int_{-0.5}^{0.5} W(f)\, e^{j2\pi f(t-s)}\, df \right\} \quad \text{(C1c)}$$

$$\tilde{Q}_f(t, s) = \frac{1}{R} \sum_{r=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(Mr - s) f(Mn - t) \cdot \sum_{d=-\infty}^{\infty} \rho(M(n - r) + (s - t) + Rd) \cdot g(s - t + Rd). \quad \text{(C1d)}$$

The indexes (t, s) in (C1) are limited to a finite interval of length L_h, for which the coefficients of the analysis window are nonzero. The summations over the indexes (r, n) are therefore finite, but the summations over d may be infinite, in which case they can be alternatively evaluated in the frequency domain. When $W(f)$ is positive on a set of nonzero measure, it follows from (C1c) (and the assumption that $W(f) \ge 0$) that the matrix $\tilde{Q}_D$ is a PD matrix. Since according to Theorem 1, $U \ge 0$, $\tilde{Q}_f$ is a PSD matrix. Therefore, $(\tilde{Q}_D + \tilde{Q}_f)$ is a PD matrix, and in particular, $(\tilde{Q}_f + \tilde{Q}_f' + \tilde{Q}_D + \tilde{Q}_D')$ is nonsingular, so that existence and uniqueness of the solution of (37) are guaranteed. All the facts mentioned in Section V in the discussion below (37) are an immediate consequence of the expressions given in (C1).

B. Proof of Theorem 2

In the proof of Theorem 2, we frequently denote by x a pair of filters $(\boldsymbol{f}, \boldsymbol{h})$ corresponding to an A/S system, and use the notation $x_{r+1} \in A_t(x_r)$ to denote the rth iteration of the iterative algorithm [where $x_r \triangleq (\boldsymbol{f}^{(r)}, \boldsymbol{h}^{(r)})$]. The assumption that both (30) and (37) possess unique solutions implies that $A_1(\cdot)$, $A_2(\cdot)$, and $A_t(\cdot)$ are point-to-point mappings.

1) Consider $\hat{U}(x_{r+1}) = \hat{U}(\boldsymbol{f}^{(r+1)}, \boldsymbol{h}^{(r+1)}) \leq \hat{U}(\boldsymbol{f}^{(r)}, \boldsymbol{h}^{(r+1)}) \leq (\hat{U}(\boldsymbol{f}^{(r)}, \boldsymbol{h}^{(r)}) = \hat{U}(x_r)$ where the two inequalities follow from the definitions of steps 3) and 2) of the algorithm and the definitions of $A_2(\cdot)$ and $A_1(\cdot)$. Equality is obtained iff $\boldsymbol{h}^{(r)} = \boldsymbol{h}^{(r+1)} = A_1(\boldsymbol{f}^{(r)})$ and $\boldsymbol{f}^{(r)} = A_2(\boldsymbol{h}^{(r+1)}) = A_2(\boldsymbol{h}^{(r)})$; thus, iff the algorithm reaches a fixed point.

2) According to the definition of Γ ($\Gamma = \{x_r;\, A_t(x_r) = \{x_r\}\}$), $(\boldsymbol{f}, \boldsymbol{h}) \in \Gamma$ if and only if $\boldsymbol{h} = A_1(\boldsymbol{f})$ and $\boldsymbol{f} = A_2(\boldsymbol{h})$, i.e., iff the A/S system defined by $(\boldsymbol{f}, \boldsymbol{h})$ satisfies both (30) and (37), which are exactly the gradient equations of the error criterion $\hat{U}$. Thus, $\Gamma = \{x;\, \nabla \hat{U}(x) = 0\}$.

3) This property follows immediately from the following.

Theorem C1 [17]: If there exists a *descent function*[4] of a point-to-point continuous mapping A with respect to a set of real vectors Γ, then every limit point of an instance of A which is contained in a compact set is in the set Γ. Providing that the following lemma holds.

Lemma 1:

a) $\hat{U}$ is a descent function of A_t with respect to the set Γ, i.e., $\hat{U}(x)$ is a continuous function, and for every $y = A_t(x)$, $\hat{U}(y) \leq \hat{U}(x)$ with equality only if $x \in \Gamma$.

b) The point-to-point mapping A_t is a continuous mapping.

c) For any initial condition x_0, the sequence $\{x_r\}_{r=0}^{\infty}$ is contained in a compact set.

Proof of the Lemma:

a) Since $\hat{U}$ is a polynomial in both $\boldsymbol{h}$ and $\boldsymbol{f}$, it is a continuous function with respect to $x = (\boldsymbol{f}, \boldsymbol{h})$. We have already shown above, in part 1) of the theorem, that for $y = A_t(x)$, $\hat{U}(y) \leq \hat{U}(x)$, with equality if and only if $x \in \Gamma$. Thus, $\hat{U}$ is a descent function of A_t with respect to Γ.

b) The point-to-point mapping $A_1(\cdot)$ is a continuous mapping, since the coefficients of the matrix $\tilde{Q}_f$ and the vector $\tilde{\boldsymbol{b}}_f$ are continuous functions of $\boldsymbol{f}$, and the determinant of the PD matrix $(\tilde{Q}_D + \tilde{Q}_D' + \tilde{Q}_f + \tilde{Q}_f')$ is bounded away from zero for every $\boldsymbol{f} \in \mathbb{R}^{L_f}$. (It is well known that the solution of a nondegenerate set of linear equations $Ax = b$ with det (A) bounded away from zero is a continuous function of the coefficients of A and $\boldsymbol{b}$ [19].) Similarly, since Q_ν is assumed to be PD, the determinant of $(Q_\nu + Q_\nu' + Q_h + Q_h')$ is bounded away from zero for every $\boldsymbol{h} \in \mathbb{R}^{L_h}$. Therefore, $A_2(\cdot)$ is a continuous mapping, and thus so is $A_t(\cdot)$.

[4] $f(x)$ is a descent function of a mapping A w.r.t. the set Γ if $f(x)$ is a continuous function, and for every $y \in A(x)$, $f(y) \leq f(x)$, with strict inequality whenever $x \notin \Gamma$.

c) Due to (28),

$$\hat{U}(x) = U(x) + \left(K_0 + (\boldsymbol{h} - \tilde{\boldsymbol{h}})'\tilde{Q}_D(\boldsymbol{h} - \tilde{\boldsymbol{h}})\right)$$
$$\geq \inf U(x) + K_0 + (\boldsymbol{h} - \tilde{\boldsymbol{h}})'\tilde{Q}_D(\boldsymbol{h} - \tilde{\boldsymbol{h}})$$

where $\tilde{\boldsymbol{h}} \triangleq (\tilde{Q}_D + \tilde{Q}_D')^{-1}\tilde{\boldsymbol{b}}_D$ and $K_0 \triangleq \int_{-0.5}^{0.5} W(f)\,|D(f) - \tilde{H}(f)|^2\, df$ in which $\tilde{H}(f)$ is the frequency response associated with $\tilde{\boldsymbol{h}}$.

Since $\hat{U}(x)$ is a descent function of the algorithm A_t, $\hat{U}(x_0) \geq \hat{U}(x_r)$ for any sequence $\{x_r\}_{r=0}^{\infty}$ starting at x_0. So let $\rho_0 \triangleq \hat{U}(x_0) - \inf U(x) - K_0 \geq (\boldsymbol{h}^{(r)} - \tilde{\boldsymbol{h}})'\tilde{Q}_D(\boldsymbol{h}^{(r)} - \tilde{\boldsymbol{h}})$. Thus, the whole sequence $\{\boldsymbol{h}^{(r)}\}_{r=0}^{\infty}$ is contained in the compact set $C(\rho_0) \triangleq \{\boldsymbol{h};\, \rho_0 \geq (\boldsymbol{h} - \tilde{\boldsymbol{h}})'\tilde{Q}_D(\boldsymbol{h} - \tilde{\boldsymbol{h}})\}$ since $\tilde{Q}_D$ is a PD matrix.

The proof of Lemma 1 is completed by the following lemma.

Lemma 2: For every finite value of ρ, the set $S(\rho) = \{(\boldsymbol{f}; \boldsymbol{h});\, \boldsymbol{f} = A_2(\boldsymbol{h}) \cap \boldsymbol{h} \in C(\rho)\}$ is contained in a compact set.

Proof: Since $C(\rho)$ is a compact set, it only remains to show that $\boldsymbol{h} \in C(\rho)$ implies that $A_2(\boldsymbol{h})$ is contained in a compact set which does depend on $\boldsymbol{h}$. Let $\lambda_m(\boldsymbol{h})$ denote the minimal eigenvalue of $(Q + Q')$ as a function of $\boldsymbol{h}$, $\|\boldsymbol{f}\|$ denotes the Euclidean norm of the vector $\boldsymbol{f} = A_2(\boldsymbol{h})$, and $\|\boldsymbol{b}(\boldsymbol{h})\|$ denotes the Euclidean norm of the vector $\boldsymbol{b}$ which appears on the right-hand side of (30) for a given value of $\boldsymbol{h}$. Then, since Q_ν is a PD matrix, $\lambda_m(\boldsymbol{h}) > \lambda_0 > 0$, and therefore, $\|\boldsymbol{f}\| \leq \|\boldsymbol{b}(\boldsymbol{h})\|/\lambda_0$. Observing (B2a), we notice that $\boldsymbol{b}$ is a linear function of $\boldsymbol{h}$; thus, on the compact set $\boldsymbol{h} \in C(\rho)$, the value of $\|\boldsymbol{b}(\boldsymbol{h})\|$ is bounded from above by $B < \infty$. Therefore, $\|\boldsymbol{f}\| \leq B/\lambda_0 < \infty$, and thus the union of $A_2(\boldsymbol{h})$ over $\boldsymbol{h} \in C(\rho)$ is contained in a ball which is a compact set.

4) Since $\hat{U}$ is a nonnegative continuous function, its infimum exists, and there exists a sequence $\{x_r\}_{r=0}^{\infty}$ starting on x_0 defined above such that $\lim_{r \to \infty} \hat{U}(x_r) = \inf_x \hat{U}(x)$. Furthermore, without loss of generality, we can assume that for any value of r, $\hat{U}(x_r) \leq \hat{U}(x_0)$. Now, if we replace every element $x_r = (\boldsymbol{f}_r, \boldsymbol{h}_r)$ of this sequence by $\tilde{x}_r = (\tilde{\boldsymbol{f}}_r, \boldsymbol{h}_r)$ with $\tilde{\boldsymbol{f}}_r \in A_2(\boldsymbol{h}_r)$, then $\inf_x \hat{U}(x) \leq \hat{U}(\tilde{x}_r) \leq \hat{U}(x_r)$, and therefore $\hat{U}(\tilde{x}_r)$ also converges to $\inf_x \hat{U}(x)$. Since $\hat{U}(\tilde{x}_r) \leq \hat{U}(x_0)$, it follows from Lemma 1, part c) that $\boldsymbol{h}_r \in C(\rho_0)$, and therefore $\tilde{x}_r \in S(\rho_0)$. Thus, the sequence $\{\tilde{x}_r\}_{r=0}^{\infty}$ is contained in a compact set, and therefore it has at least one limit point there. Let x^* denote a limit point, and let $\tilde{x}_{r_k}$ be the subsequence that converges to x^*. Since $\tilde{x}_{r_k} \underset{k \to \infty}{\to} x^*$, $\hat{U}(\tilde{x}_{r_k}) \underset{k \to \infty}{\to} \inf_x \hat{U}(x)$, and $\hat{U}(x)$ is a continuous function, $\hat{U}(x^*) = \inf_x \hat{U}(x)$, i.e., there exists a global minimum of $\hat{U}(x)$ which is in Γ since it is a stationary point of $\hat{U}(x)$.

Acknowledgment

The authors thank Z. Sela of the Department of Mathematics, Technion—Israel Institute of Technology, for the helpful discussion which led to the statement and proof of Theorem 1. The assistance of A. Satt who programmed the design examples and performed the computer simulations with speech signals is gratefully acknowledged. Many suggestions of an anonymous reviewer which improved the presentation of this work are appreciated.

REFERENCES

[1] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall Signal Processing Series, 1983, ch. 7.

[2] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, Oct. 1979.

[3] L. S. Lee, G. C. Chan, and C. S. Chang, "A new frequency domain speech scrambling system which does not require frame synchronization," *IEEE Trans. Commun.*, vol. COM-32, pp. 444–457, Apr. 1984.

[4] V. K. Jain and R. E. Crochiere, "A novel approach to the design of analysis/synthesis filter banks," in *Proc. ICASSP'83*, Boston, MA, pp. 228–231.

[5] Z. Shpiro and D. Malah, "An algebraic approach to discrete short time Fourier transform analysis and synthesis," in *Proc. ICASSP'84*, pp. 2.3.1.–2.3.4.

[6] D. W. Griffin and J. S. Lim, "Signal estimation from modified short time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 236–243, Apr. 1984.

[7] H. Abut and S. A. Luse, "Vector quantizers for sub-band coded waveforms," in *Proc. ICASSP'84*, pp. 10.6.1–10.6.4.

[8] R. E. Crochiere, "On the design of sub-band coders for low bit rate speech communication," *Bell Syst. Tech. J.*, pp. 741–771, May–July 1977.

[9] ——, "A weighted overlap-add method of short time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99–102, Feb. 1980.

[10] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978, ch. 6.

[11] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984, sec. 4.7.

[12] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55–69, Feb. 1980.

[13] Z. Shpiro, "Analog speech scrambling by means of discrete short-time Fourier transform," M.Sc. thesis (in Hebrew), Technion, I.I.T., Haifa, Nov. 1983.

[14] A. Dembo and D. Malah, "Statistical design of analysis/synthesis systems with quantization," EE. Publ. 585, Technion, I.I.T., Haifa, Apr. 1986.

[15] P. J. Davis, *Circulant Matrices*. New York: Wiley, 1979, ch. 2, 3.

[16] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall, 1974, ch. 18, 19.

[17] D. G. Luenberger, *Introduction to Linear and Non-Linear Programming*. Reading, MA: Addison-Wesley, 1973, sec. 6.5.

[18] Y. Katzenelson, *An Introduction to Harmonic Analysis*. New York: Wiley, 1968, ch. 1.

[19] P. Lancaster, *Theory of Matrices*. New York: Academic, 1969, ch. 3.

[20] B. H. Juang and A. H. Gray, Jr., "Multiple stage vector quantization for speech coding," in *Proc. ICASSP'82*, Paris, France, pp. 597–600.

[21] J.-H. Chen and A. Gersho, "Gain-adaptive vector quantization for medium-rate speech coding," in *Proc. IEEE Int. Conf. Commun.*, 1985, pp. 1456–1460.

Amir Dembo (S'84–M'86), for a photograph and biography, see p. 181 of the February 1988 issue of this TRANSACTIONS.

David Malah (S'67–M'71–SM'84–F'87), for a photograph and biography, see p. 181 of the February 1988 issue of this TRANSACTIONS.