Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial

P. P. VAIDYANATHAN, SENIOR MEMBER, IEEE

Multirate digital filters and filter banks find application in communications, speech processing, image compression, antenna systems, analog voice privacy systems, and in the digital audio industry. During the last several years there has been substantial progress in multirate system research. This includes design of decimation and interpolation filters, analysis/synthesis filter banks (also called quadrature mirror filters, or QMF), and the development of new sampling theorems. First, the basic concepts and building blocks in multirate digital signal processing (DSP), including the digital polyphase representation, are reviewed. Next, recent progress as reported by several authors in this area is discussed. Several applications are described, including the following: subband coding of waveforms, voice privacy systems, integral and fractional sampling rate conversion (such as in digital audio), digital crossover networks, and multirate coding of narrow-band filter coefficients. The M-band QMF bank is discussed in considerable detail, including an analysis of various errors and imperfections. Recent techniques for perfect signal reconstruction in such systems are reviewed. The connection between QMF banks and other related topics, such as block digital filtering and periodically time-varying systems, based on a pseudo-circulant matrix framework, is covered. Unconventional applications of the polyphase concept are discussed.

I. INTRODUCTION

In recent years there has been tremendous progress in the multirate processing of digital signals. Unlike the single-rate system, the sample spacing in a multirate system can vary from point to point [1], [2]. This often results in more efficient processing of signals because the sampling rates at various internal points can be kept as small as possible. Unfortunately, this also results in the introduction of a new type of error, i.e., aliasing, which should somehow be canceled eventually.

The basic building blocks in a multirate digital signal processing (DSP) system are decimators and interpolators. In 1981, an excellent tutorial article on decimation and interpolation appeared in [3]. Subsequent to this a text on the subject of multirate systems has also been published by the same authors [4]. Since then, a number of new developments have taken place in the area, particularly in multirate

digital filter bank designs. A short summary of some of these developments was reported recently by this author at an IEEE international conference [5]. The purpose of this article is to provide a self-contained and more complete exposure to many recent contributions on multirate systems, including filter bank design.

As mentioned in [3], multirate systems find application in communications, speech processing, spectrum analysis [6], radar systems, and antenna systems. In this tutorial, two sections are devoted to a review of applications. In Section III, we point out applications in digital audio systems, in subband coding techniques (used in speech and image compression), and in analog voice privacy systems (for standard telephone communications). In Section V-E, applications of special transfer functions (such as complementary functions) in digital audio is reviewed. In Section IX, several unconventional applications of multirate systems and polyphase theory are indicated. These include a) derivation of new sampling theorems for efficient compression of signals, b) derivation of new techniques for efficient coding of impulse response sequences of narrow band filters, c) design of FIR filters with adjustable multilevel responses, and d) adaptive filtering in subbands.

A. Paper Outline

In section II, basic tools, such as decimators, interpolators, decimation and interpolation filters, and digital filter banks, are reviewed, along with the interconnection properties of the building blocks. In section III, some applications of multirate DSP are indicated, in digital audio systems, in subband coding, and in voice privacy systems. Section IV reviews the digital polyphase decomposition due to Bellanger, along with applications such as the uniform DFT filter bank. The concept of multilevel polyphase decomposition is also introduced here as a tool for efficient implementation of fractional decimation filters. Several special types of filter banks, such as Nyquist filters, power-complementary systems and Euclidean filter-banks, are studied in section V. In section VI, the two-band QMF bank is studied in sufficient detail along with procedures for eliminating aliasing in such systems. Procedures for elimination of amplitude and/or phase distortion are discussed.

Perfect-reconstruction two-channel QMF banks are introduced by blending the polyphase concept with the classical network-theoretic concept of losslessness.

The relation between M-band QMF banks and two other related topics (block filtering and periodically time-varying systems) is reviewed in section VII, based on an algebraic structure called the pseudo-circulant matrix. In section VIII, M-band QMF banks are discussed in greater detail, and techniques for elimination of aliasing, amplitude, and phase distortions are reviewed. Section IX discusses unconventional applications, and Section X discusses some extensions of multirate ideas to cases of multidimensional signals. The paper concludes with a discussion of open problems in multirate DSP.

B. Notations and Terminology

The variables Ω and ω are used as frequency variables for continuous-time and discrete-time cases, respectively. In the discrete-time case the term normalized frequency is used to denote $f = \omega/2\pi$. The frequency response of a transfer function $H(z)$ is expressed as $H(e^{j\omega}) = |H(e^{j\omega})| e^{j\phi(\omega)}$, where $|H(e^{j\omega})|$ is the magnitude response and $\phi(\omega)$ the phase response. The quantity $\tau(\omega) = -d\phi(\omega)/d\omega$ is the group delay of $H(z)$. If $|H(e^{j\omega})|$ is constant for all ω, $H(z)$ is all-pass. If $\phi(\omega)$ has the form $k_0 - k_1\omega$, then $H(z)$ is said to have linear phase and the group delay is a constant k_1; physically, if the input to such a filter $H(z)$ has energy only in the passband of $H(z)$, then the output is a delayed version of the input, by k_1 samples. Unless mentioned otherwise, a low-pass filter has real coefficients, so that $|H(e^{j\omega})|$ is symmetric and $\phi(\omega)$ is antisymmetric with respect to $\omega = 0$. Usually $|H(e^{j\omega})|$ is plotted for $0 \le f \le 0.5$ (i.e., for $0 \le \omega \le \pi$). If ω_p and ω_s denote the passband and stopband edges of a low-pass filter, the quantity $\omega_c = (\omega_p + \omega_s)/2$ is said to be the cutoff frequency.

Bold-faced quantities denote matrices and vectors, as in A, $H(z)$ etc. The symbol I_k denotes the $k \times k$ identity matrix (with subscript often omitted). The quantities A^T, $A^\dagger$ and A^* denote, respectively, the transpose, transpose conjugate, and conjugate of A. For functions $H(z)$, the notation $H_*(z)$ denotes conjugation of the coefficients without conjugating z. For example if $H(z) = a + bz^{-1}$, then $H_*(z) = a^* + b^*z^{-1}$. Thus, $H^*(z) = H_*(z^*)$. The notation $\tilde{H}(z)$ stands for $H_*^T(z^{-1})$. In other words, conjugate the coefficients, take transpose (if matrix), and replace z with z^{-1}. When $z = e^{j\omega}$ (i.e., on the unit circle), we have $\tilde{H}(z) = H^\dagger(z)$. Linear time-invariant systems [7] are abbreviated as LTI and linear periodically time-varying systems as LPTV. A $p \times r$ matrix A is said to be unitary (orthogonal if it is real) if $A^\dagger A = cI_r$, $c \ne 0$. Note that A is not restricted to be square. For example, $[\begin{smallmatrix} \cos(\theta) \\ \sin(\theta) \end{smallmatrix}]$ is unitary for any real θ. The symbol W_M stands for $e^{-j2\pi/M}$. The subscript M is usually deleted because its value is often clear from the context. This quantity appears in the definition of the discrete Fourier transform (DFT) [7], [8]. Thus an M-point sequence $[x_0, x_1, \cdots, x_{M-1}]$ has the M-point DFT sequence

$$X_k = \sum_{n=0}^{M-1} x_n W^{kn}, \qquad 0 \le k \le M - 1. \tag{1a}$$

The inverse DFT (IDFT) is given by

$$x_n = \frac{1}{M} \sum_{k=0}^{M-1} X_k W^{-nk}. \tag{1b}$$

The most crucial property of W that finds repeated use in multirate DSP is the following:

$$\sum_{n=0}^{M-1} W^{kn} = \begin{cases} M, & k = \text{integer multiple of } M \\ 0, & \text{otherwise.} \end{cases} \tag{1c}$$

For any pair of integers k, n, we have $W^k = W^n$, if and only if $k - n$ is an integer multiple of M. In particular, therefore, $W^k \ne W^n$ for $0 \le k < n \le M - 1$.

State Space Descriptions: Consider a discrete time transfer matrix $H(z)$ with input vector $u(n)$ and output vector $y(n)$. Suppose we have implemented this transfer matrix using a structure, and let N denote the number of delay elements used. Label the outputs of the delay elements as the state variables $x_k(n)$, $0 \le k \le N - 1$, and define the state vector $x(n) = [x_0(n) \, x_1(n) \cdots x_{N-1}(n)]^T$. With $u(n)$ and $y(n)$ denoting the input and output (vector) sequences to the structure, one can always find equations of the form [9]

$$x(n + 1) = Ax(n) + Bu(n),$$

$$y(n) = Cx(n) + Du(n),$$

to describe the structure. This is called the state–space description of the structure. The matrix A, called the state transition matrix, has size $N \times N$, where N is the number of delays in the structure. The transfer function $H(z)$ of the structure is given by $H(z) = D + C(zI - A)^{-1}B$. The smallest number of delay elements (i.e., z^{-1} elements) required to implement $H(z)$ is called the McMillan degree (or simply, the degree) of $H(z)$. If the number of delays N in the structure is equal to the degree, then the structure is said to be a minimal realization of $H(z)$. This is equivalent to saying that A is as small as possible.

A summary of acronyms and common notations used in this paper is found in the Nomenclature, which follows section XI.

II. Basic Building Blocks and Tools

In this section we introduce the basic multirate building blocks, along with their frequency-domain characterizations, and interconnection behaviors.

A. Decimators and Interpolators

Fig. 1 shows block diagrams of these building blocks. The decimator is characterized by the input–output relation
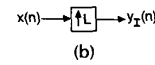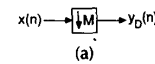
$$y_D(n) = x(Mn) \tag{2a}$$

(a)

(b)

Fig. 1. Building blocks. (a) M-fold decimator. (b) L-fold interpolator.

which says that the output at time n is equal to the input at time Mn. As a consequence, only the input samples with sample numbers equal to a multiple of M are retained. This sampling-rate reduction by a factor of M is demonstrated in Fig. 2 for the case of $M = 2$. The L-fold interpolator is char-
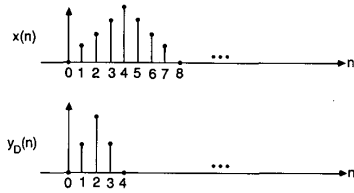
Fig. 2. Demonstration of decimation for $M = 2$.

acterized by the input–output relation

$$y_I(n) = \begin{cases} x\left(\dfrac{n}{L}\right) & \text{if } n \text{ is a multiple of } L \\ 0 & \text{otherwise.} \end{cases} \quad (2b)$$

That is, the output $y_I(n)$ is obtained by inserting $L - 1$ zero-valued samples between adjacent samples of $x(n)$, as demonstrated in Fig. 3 for $L = 2$. The decimator and interpolator
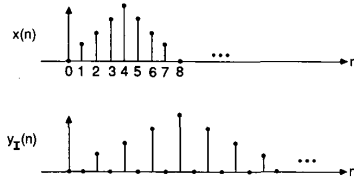

Fig. 3. Demonstration of interpolation for $L = 2$.

are linear systems even though they are time-varying [4], [5], [10].

The z transform of the interpolator output $y_I(n)$ is given by [4]:

$$Y_I(z) = X(z^L). \quad (3a)$$

This means $Y_I(e^{j\omega}) = X(e^{j\omega L})$ i.e., $Y_I(e^{j\omega})$ is an L-fold compressed version of $X(e^{j\omega})$, as demonstrated in Fig. 4(b). The appearance of multiple copies of the basic spectrum in Fig. 4 is called the imaging effect and the extra copies are the images created by the interpolator.
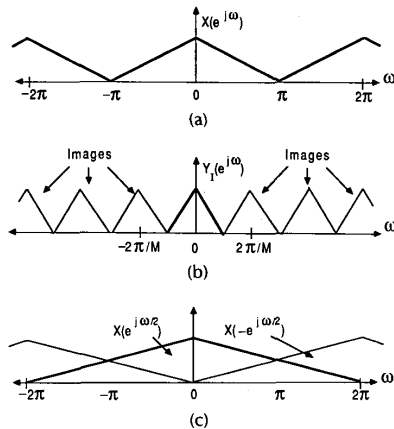

Fig. 4. Transform-domain effects of decimation and interpolation. (a) The z transform. (b) L-fold compressed version. (c) Demonstration of effect when $M = 2$.

Since decimation corresponds to compression in the time domain, one might expect a stretching effect in the frequency domain. To be more precise, the z transform of $y_D(n)$ is given by

$$Y_D(z) = \frac{1}{M} \sum_{k=0}^{M-1} X(z^{1/M} W^k) \quad (3b)$$

which means $MY_D(e^{j\omega}) = \sum_{k=0}^{M-1} X(e^{j(\omega - 2\pi k)/M})$. The term with $k = 0$ is indeed the M-fold stretched version of $X(e^{j\omega})$. The $M - 1$ terms with $k > 0$ are uniformly shifted versions of this stretched version. These M terms together make up a function with period 2π in ω, which is the basic property of the Fourier transform of any sequence [8]. Fig. 4(c) demonstrates this effect for $M = 2$. The terms with $k > 0$ are called the aliasing terms. As long as $x(n)$ is bandlimited to $|\omega| < \pi/M$, there is no overlap of these terms with the $k = 0$ term.

The fundamental difference between aliasing and imaging is important to notice. Aliasing can cause loss of information because of the possible overlap of the shifted versions of the stretched version of $X(e^{j\omega})$. Imaging, on the other hand, does not lead to any loss of information (which is consistent with the fact that no time-domain samples are lost).

B. Interconnections

Fig. 5 shows a cascade connection which is often encountered in filter-bank systems. The signal $v(n)$ here is equal to
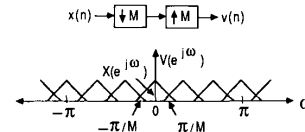

Fig. 5. Effect of decimation followed by interpolation.

$x(n)$ whenever n is a multiple of M, and zero otherwise. The transform-domain relation is

$$V(z) = \frac{1}{M} \sum_{k=0}^{M-1} X(zW^k) \quad (4)$$

which means that $MV(e^{j\omega})$ is a sum of $X(e^{j\omega})$ with the $M - 1$ uniformly shifted versions $X(e^{j(\omega - 2\pi k/M)})$. From the figure we see that $x(n)$ can be recovered from $v(n)$ by eliminating the images by filtering, provided none of the images has an overlap with $X(e^{j\omega})$. If such an overlap occurs, it implies aliasing and $x(n)$ cannot be recovered. Notice that in order for $x(n)$ to be recoverable it is not necessary for $X(e^{j\omega})$ to be restricted to $|\omega| < \pi/M$. It is sufficient for the total bandwidth of $X(e^{j\omega})$ to be less than $2\pi/M$. Thus a general bandpass signal with energy in the region $a \le \omega \le a + 2\pi/M$ can be decimated by M without creating overlap of the alias components, and the decimated signal in general is a full-band signal.

A different type of cascade is shown in Fig. 6(a). We shall have occasion to use this in section IV-B, which concentrates on multilevel-polyphase decompositions. It should be cautioned that the two building blocks in Fig. 6(a) are not, in general, interchangeable, i.e., the systems in Fig. 6(a) and 6(b) are not equivalent. For example, with $M = L$, the system of Fig. 6(a) is an identity system, whereas the system of Fig. 6(b) causes a loss of $M - 1$ out of M samples. It can be shown
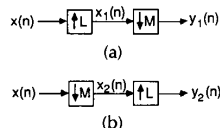
Fig. 6. Two ways to cascade decimator and interpolator. These are equivalent if and only if M and L are relatively prime. (a) Example of identity system. (b) Example of loss of $M - 1$ out of M samples.

(Appendix A) that the systems of Fig. 6(a) and 6(b) are identical if and only if L and M are relatively prime.

Decimation Filters and Interpolation Filters: In most applications a decimator is preceded by a bandlimiting filter $H(z)$ whose purpose is to avoid aliasing. For example, a low-pass filter with stopband edge $\omega_s = \pi/M$ can serve as such a filter. The cascade shown in Fig. 7(a) is commonly called a decimation filter. An interpolation filter, on the other hand, is a device which follows an interpolator (Fig. 7(b)), the purpose being to eliminate the images. The low-pass filter of Fig. 7(c) again serves as an example (with $L = M$).
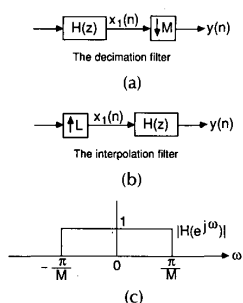

Fig. 7. (a) Decimation filter. (b) Interpolation filter. (c) Low-pass filter.

Fractional Sampling Rate Alterations: Fig. 8(a) shows a scheme for reducing the sampling rate by a nonintegral (rational) number M/L. Fractional reduction of sampling rate often results in data compression without loss of information. As an example, if $X(e^{j\omega})$ is as in Fig. 8(b), then a fractional reduction by 3/2 is possible. This can be accom-
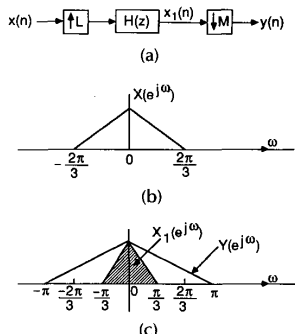

Fig. 8. Decimation by rational fraction of M/L. (a) General structure. (b) Example of bandlimited signal. (c) Effect of fractional decimation of this signal ($L = 2$, $M = 3$).

plished by setting $M = 3$, $L = 2$ in Fig. 8(a). The filter $H(z)$ is then taken to be low pass, with passband edge at $\pi/3$ and stopband edge at $2\pi/3$. Notice that in this application, the transition bandwidth of $H(z)$ need not be unduly narrow. The various signals in Fig. 8(a) have transforms as in Fig. 8(c), so that $Y(e^{j\omega})$ is a fractionally stretched version of $X(e^{j\omega})$.

Two Noble Identities: In Fig. 9(a) we have a decimator followed by a transfer function $G(z)$. It can be proved, based
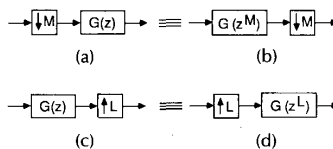

Fig. 9. Noble identities for multirate systems. (a) Decimator followed by transfer function $G(z)$. (b) Equivalent cascade. (c) Example of transfer function preceding. (d) Equivalent cascade.

on (3b), that this cascade is equivalent to the one in Fig. 9(b) provided $G(z)$ is a rational transfer function (i.e., a ratio of polynomials in z^{-1}). In a similar manner, the two cascades in Figs. 9(c) and 9(d) are equivalent (provided $G(z)$ is rational), as can be proved from (3a). These identities are very valuable in practically all applications for efficient implementation of filters and filter banks. We shall call these the "noble identities."

C. Analysis and Synthesis Banks

These are the two basic types of filter banks. An analysis bank is a set of analysis filters $H_k(z)$ which splits a signal into M subband signals $x_k(n)$ as shown in Fig. 10(a). What we do
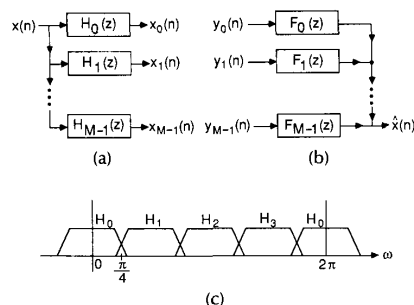

Fig. 10. Analysis and synthesis filter banks. (a) Analysis bank. (b) Synthesis bank. (c) Typical response of uniform DFT filter bank; here $M = 4$.

with the subband signals depends on the application, as we shall see in sections III, VI, and IX. Next, a synthesis bank (Fig. 10(b)) consists of M synthesis filters $F_k(z)$, which combine M signals $y_k(n)$ (possibly from an analysis bank) into a reconstructed signal $\hat{x}(n)$. There are several types of filter banks, i.e., the complementary type, the Nyquist type, etc., to be described in Section V along with applications.

Uniform DFT Filter Banks: An analysis bank with M filters ($M > 1$) is said to be a uniform DFT filter bank if all the filters are derived from $H_0(z)$ according to $H_k(z) = H_0(zW^k)$, $0 \leq k \leq M - 1$. Here $H_0(z)$ is called the prototype filter. Note

that $H_k(e^{j\omega}) = H_0(e^{j(\omega - 2\pi k/M)})$, which means that the frequency responses of $H_k(z)$ are uniformly shifted versions of $H_0(e^{j\omega})$. Fig. 10(c) shows a typical set of responses, where $H_0(z)$ is taken to be low pass. More details can be found in section IV-C and in [4] and [11].

III. SOME APPLICATIONS OF MULTIRATE SYSTEMS

We shall now review a number of important applications of multirate filters and filter banks, with pointers to the literature for details, examples, and demonstrations. In section IX, several unconventional applications are also outlined.

Applications in the design of transmultiplexers (which are devices for conversion between frequency division multiplexing (FDM) and time-division multiplexing (TDM)) are not discussed here in detail, primarily because of the excellent treatment already available in [13]. Also see [14] for the correspondence between transmultiplexers and analysis/synthesis filter banks. The input to a TDM-to-FDM converter is a signal $y(n)$, which is the time-multiplexed version of M signals $y_k(n)$, $0 \le k \le M - 1$. Given $y(n)$, the components $y_k(n)$ can easily be separated out by use of a commutator switch [4], [13]. These M signals are then modulated using distinct carrier frequencies. The carrier frequencies ω_k, $0 \le k \le M - 1$ are chosen so that there is sufficient spectral gap between the messages. A sum of these M signals (which is the FDM signal) is then transmitted through the channel. The total channel bandwidth is therefore required to exceed the sum of signal bandwidths because of the safeguard gap between adjacent spectra. The gap enables one to obtain perfect recovery of the multiplexed signals $y_k(n)$ at a future point.

A novel approach to transmultiplexing was suggested in [36] and cited in [14], based on synthesis and analysis filter banks. This approach permits overlap between the spectra of successive messages in the frequency domain. The total required channel bandwidth is therefore less than that in conventional FDM channels. Conditions are derived under which cross-talk can be avoided and the set of M original signals can still be recovered from this version. Details can be found in reference [36] cited in [14].

A. Digital Audio Systems

In the digital audio industry, it is a common requirement to change the sampling rates of band-limited sequences. This arises for example when an analog music waveform $x_a(t)$ is to be digitized. Assuming that the significant information is in the band $0 \le |\Omega|/2\pi \le 22$ kHz [15], a minimum sampling rate of 44 kHz is suggested (Fig. 11(a)). It is, however, necessary to perform analog filtering before sampling to eliminate aliasing of out-of-band noise. Now the requirements on the analog filter $H_a(j\Omega)$ (Fig. 11(b)) are stringent: it should have a fairly flat passband (so that $X_a(j\Omega)$ is not distorted) and a narrow transition band (so that only a small amount of unwanted energy is let in). Optimal filters for this purpose (such as elliptic filters [9], which are optimal in the minimax sense) have a very nonlinear phase response [16, page 82] around the bandedge (i.e., around 22 kHz). In high-quality music this is considered to be objectionable [15]. A common strategy to solve this problem is to oversample $x_a(t)$ by a factor of two (and often four). The filter $H_a(j\Omega)$ now has a much wider transition band, as in Fig. 11(c), so that
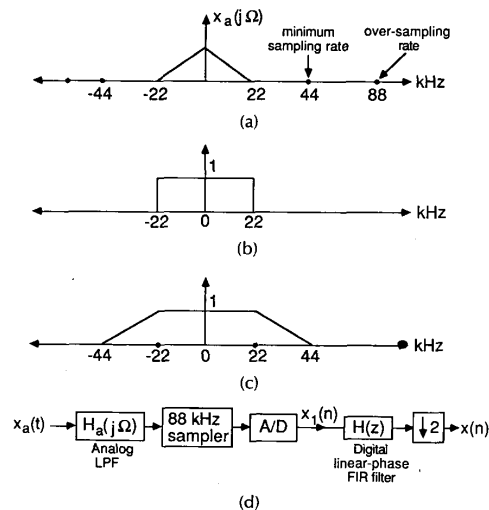

Fig. 11. (a) Spectrum of $x_a(t)$. (b) Antialiasing filter response for sampling at 44 kHz. (c) Antialiasing filter response for sampling at 88 kHz. (d) Improved scheme for A/D stage of digital audio system.

the phase-response nonlinearity is acceptably low. A simple analog Bessel filter (which has linear phase in the passband [9]) can be used in practice. The sequence $x_1(n)$ so generated is then lowpass filtered (Fig. 11(d)) by a digital filter $H(z)$ and then decimated by the same factor of two to obtain the final digital signal $x(n)$. The crucial point is that since $H(z)$ is digital, it can be designed to have linear phase [7], [16], [17], while at the same time providing the desired degree of sharpness.

A similar problem arises after the D/A conversion stage, where the digital music signal $y(n)$ should be converted to an analog signal by lowpass filtering. To eliminate the images of $Y(e^{j\omega})$ in the region outside 22 kHz, a sharp cutoff (hence nonlinear phase) analog low-pass filter is required. This problem is avoided by using an interpolation filter, as in Fig. 7(b), which increases the sampling rate digitally. After this, D/A conversion is performed followed by analog filtering. The interpolation filter $H(z)$ is once again a linear-phase FIR low-pass filter and introduces no phase distortion.

The obvious price paid in these systems is the increased internal rate of computation. However, by using the polyphase framework (section IV) the efficiency of these multirate systems can be dramatically improved.

In digital audio, it is relatively economic (compared to the analog case) to produce high-quality copies of material from one medium to another [15]. Perhaps to discourage such practice, the sampling rates used for various media are often made different from each other. It is therefore necessary in studios to design efficient nonintegral sampling rate converters (such as the one in Fig. 8(a)). See section IV-B for further details. Further applications of multirate filter banks in digital audio can be found in section V-E.

B. Subband Coding of Speech and Image Signals

In practice, one often encounters signals with energy dominantly concentrated in a particular region of fre-

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.

fastcase
Smarter legal research.