A Unified Approach to Short-Time Fourier Analysis and Synthesis

JONT B. ALLEN AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—Two distinct methods for synthesizing a signal from its short-time Fourier transform have previously been proposed. We call these methods the filter-bank summation (FBS) method and the overlap add (OLA) method. Each of these synthesis techniques has unique advantages and disadvantages in various applications due to the way in which the signal is reconstructed. In this paper we unify the ideas behind the two synthesis techniques and discuss the similarities and differences between these methods. In particular, we explicitly show the effects of modifications made to the short-time transform (both fixed and time-varying modifications are considered) on the resulting signal and discuss applications where each of the techniques would be most useful. The interesting case of nonlinear modifications (possibly signal dependent) to the short-time Fourier transform is also discussed. Finally it is shown that a formal duality exists between the two synthesis methods based on the properties of the window used for obtaining the short-time Fourier transform.

I. Introduction

THE CONCEPTS of short-time Fourier analysis and synthesis are fundamental for describing any quasi-stationary (slowly time varying) signal such as speech. With the advent of the fast Fourier transform, as well as modern digital filtering techniques, implementations of signal processing systems based on the short-time Fourier transform have become practical and are used in many applications [1]–[4]. The theory behind short-time Fourier analysis and synthesis has evolved in several discrete and usually disconnected steps [1]–[8]. It is the purpose of this paper to unify the various methods of analysis and synthesis, and to show the effects of modifying the short-time Fourier transform on the resulting signal.

II. Definition of the Short-Time Fourier Transform

Let $x(n)$ be a signal[1] defined for all n, and let $X_n(e^{j\omega_k})$ be the short-time Fourier transform of $x(n)$ evaluated at time n and frequency ω_k. In general one can define the short-time Fourier transform in terms of the output of an arbitrary bank of filters. However, we shall restrict ourselves to the much simpler case of identical, symmetric, bandpass filters uniformly spaced in frequency. The result of these simplifications is to allow the use of a single low-pass filter (window function) $w(n)$ which determines all of the properties of the filter bank. The short-time Fourier transform may then be defined as [4]

$$X_n(e^{j\omega_k}) = \sum_{m=-\infty}^{\infty} w(n-m)\, x(m)\, e^{-j\omega_k m}. \tag{2.1}$$

Equation (2.1) shows that $w(n)$, the window, selectively determines the portion of $x(n)$ which is being analyzed.

Manuscript received March 1, 1977; revised May 24, 1977.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

[1] Our results may be equally well stated in a continuous time-domain formulation.

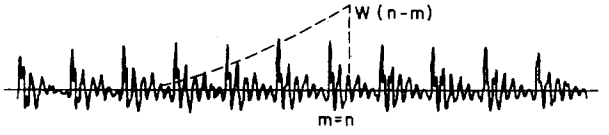

Fig. 1. An interpretation of the weighting of the signal for computing the short-time spectrum.

Fig. 1 shows a typical example of the signals involved in the computation of (2.1) for $w(n)$ an exponentially decaying window. For this window it is seen that the analysis weights the most recent samples (i.e., values of $x(m)$ near $m = n$) most heavily in computing the short-time Fourier transform.

Two equivalent but distinct interpretations may be given to (2.1). The first interpretation is that of a filter-bank analysis in which $X_n(e^{j\omega_k})$ is viewed as a function of n for a fixed ω_k. In this case $X_n(e^{j\omega_k})$ can be written as the linear convolution (denoted here by $*$) of the signal $x(n)\, e^{j\omega_k n}$ with the impulse response $w(n)$, i.e.,

$$X_n(e^{j\omega_k}) = [x(n)\, e^{-j\omega_k n}] * w(n) \tag{2.2}$$

where $w(n)$ is a low-pass filter being applied to the signal $x(n)\, e^{-j\omega_k n}$. The modulation of $x(n)$ by $e^{-j\omega_k n}$ serves to shift the frequency spectrum of $x(n)$ at frequency ω_k to 0 frequency. Thus the short-time Fourier transform can be thought of as filtering the shifted spectrum of $x(n)$ in the region of frequency ω_k by the low-pass filter $w(n)$.

The second interpretation of $X_n(e^{j\omega_k})$ is as the normal Fourier transform (i.e., z-transform evaluated on the unit circle) of the modified sequence

$$y_n(m) = x(m)\, w(n-m). \tag{2.3}$$

For this case we interpret $X_n(e^{j\omega_k})$ as a function of ω_k for a fixed value of n. Equation (2.3) shows that, for n constant, $y_n(m)$ is a product of x and w. Thus the normal Fourier transform of y_n is the complex convolution of the Fourier transforms of x and w. As such the details of the resulting short-time Fourier transform are greatly influenced by the choice of windows. Thus it is important to design a window consistent with the desired time and frequency resolution of the short-time transform. By way of example, assume the window is causal, and of duration N samples, i.e.,

$$\begin{aligned} w(n) &= 0, \quad n < 0, n > N \\ &\neq 0, \quad 0 \leqslant n \leqslant N-1. \end{aligned} \tag{2.4}$$

Fig. 2 shows plots of typical short-time transforms of voiced speech for $w(n)$ a Hamming window—i.e.,

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/N), & 0 \leqslant n \leqslant N-1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

Part (a) of this figure shows the signal $y_n(m)$ of (2.3) for

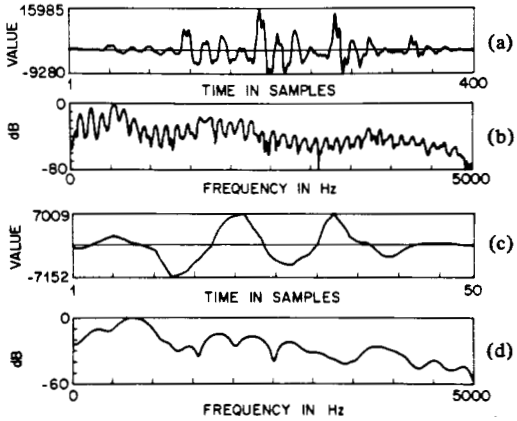

Fig. 2. Typical signals and the resulting log magnitude spectra for $w(n)$ a Hamming window of length $N = 400$ (parts (a) and (b)) and $N = 50$ (parts (c) and (d)).

$N = 400$, and part (b) shows the log magnitude (in dB) of $X_n(e^{j\omega_k})$ evaluated at the set of frequencies

$$\omega_k = \frac{2\pi k}{L}, \qquad k = 0, 1, \cdots, \frac{L}{2} \tag{2.6}$$

where $L = 1024$ in this case. Similarly Figs. 2(c) and 2(d) show $y(n)$ and the resulting log magnitude short-time spectrum for $N = 50$. (The reader should note the different time scales for parts (a) and (c) of this figure.) The differences in resolution, and level of detail of the resulting spectra are apparent.

For some applications (e.g., spectrum displays [9], signal detection, speech pitch [10], and formant estimation [11]) the short-time Fourier transform is used primarily as a nonstationary representation of the signal properties. In such cases no synthesis procedure is required. However, for many other applications, the short-time Fourier transform is used as an *intermediate* representation of the signal. Examples of these applications include vocoders [1], [4] and signal processors where we wish to modify the short-time transform in a way that may take advantage of nonstationary spectral properties of the signal [2], [3], [12]. As such, a method for reconstructing the signal $x(n)$ from its short-time transform is required. In the next sections, we discuss the two known synthesis methods [4], [8].

III. Filter-Bank Summation (FBS) for Short-Time Synthesis

The first method of reconstructing $x(n)$ is the classical method which is related to the filter-bank interpretation of the short-time Fourier transform. It was shown in Section II that for any frequency ω_k, $X_n(e^{j\omega_k})$ is a low-pass representation of the signal in a band centered at ω_k. Thus a reasonable synthesis method is to modulate $X_n(e^{j\omega_k})$ back to frequency ω_k, and then sum the result over frequency. The first operation results in the signal

$$y_k(n) = X_n(e^{j\omega_k})\, e^{j\omega_k n}. \tag{3.1}$$

We then obtain the reconstructed signal $y(n)$ as

$$y(n) = \sum_k y_k(n) \tag{3.2}$$

$$= \sum_k X_n(e^{j\omega_k})\, e^{j\omega_k n} \tag{3.3}$$

where the sum over k extends over the number of frequencies used in the analysis.

To show that $y(n)$ equals the original signal $x(n)$, we use (2.1) in (3.3) to give

$$y(n) = \sum_k \left[\sum_m w(n-m)\, x(m)\, e^{-j\omega_k m} \right] e^{j\omega_k n}. \tag{3.4}$$

Interchanging orders of summation gives

$$y(n) = \sum_m w(n-m)\, x(m) \sum_k e^{j\omega_k (n-m)}. \tag{3.5}$$

If we assume that the analysis is performed at L uniformly spaced frequencies (as in (2.6)) we can sum over k in (3.5) giving

$$y(n) = \sum_m w(n-m)\, x(m) \sum_{r=-\infty}^{\infty} L\delta(n-m-rL) \tag{3.6}$$

where $\delta(n) = 1$ for $n = 0$ and is zero for $n \neq 0$. Evaluating (3.6) for $m = n - rL$ (i.e., when $\delta(n-m-rL) = 1$) gives

$$y(n) = L \sum_{r=-\infty}^{\infty} w(rL)\, x(n-rL). \tag{3.7}$$

Since $w(n)$ is of duration N samples, we see that if $L \geq N$ then (3.7) can be truncated to the $r = 0$ term giving

$$y(n) = Lw(0)\, x(n). \tag{3.8}$$

Thus for $L \geq N$ the reconstructed sample y at time n is a scaled (by $Lw(0)$) replica of the input sample x—i.e., *the short-time Fourier transform representation is exactly invertible by the FBS method.* In Section VIII, we will show that this synthesis procedure is based on the identity

$$Lw(0)\,\delta(n) = w(n) \sum_k e^{j\omega_k n} \tag{3.9}$$

which is always true for sufficiently dense samples of ω_k.

If L, the number of uniformly spaced analysis frequencies $(\omega_0, \omega_1, \cdots, \omega_{L-1})$, is less than the window duration N then (3.7) says that $y(n)$ cannot be *exactly* a replica of $x(n)$ unless the window satisfies the further property that

$$w(rL) = 0, \qquad r = \pm 1, \pm 2, \cdots. \tag{3.10}$$

Note that in this case (3.9) still holds. Techniques for designing windows (low-pass filters) which approximately satisfy (3.10) are given in [5] and [6]. In the remainder of the discussion we will assume that $L \geq N$ so that (3.8) holds, even when (3.10) does not. A discussion of the required "sampling rates" of $X_n(e^{j\omega_k})$ in time (n) and frequency (k) will be given in Section VII of this paper.

IV. Overlap Addition (OLA) Method for Short-Time Synthesis

An alternative method of synthesis is based on the normal Fourier transform interpretation of the short-time transform. Since $X_n(e^{j\omega_k})$ of (2.1) can be considered to be the Fourier transform of the sequence

$$\hat{x}_n(m) = x(m)\, w(n-m) \tag{4.1}$$

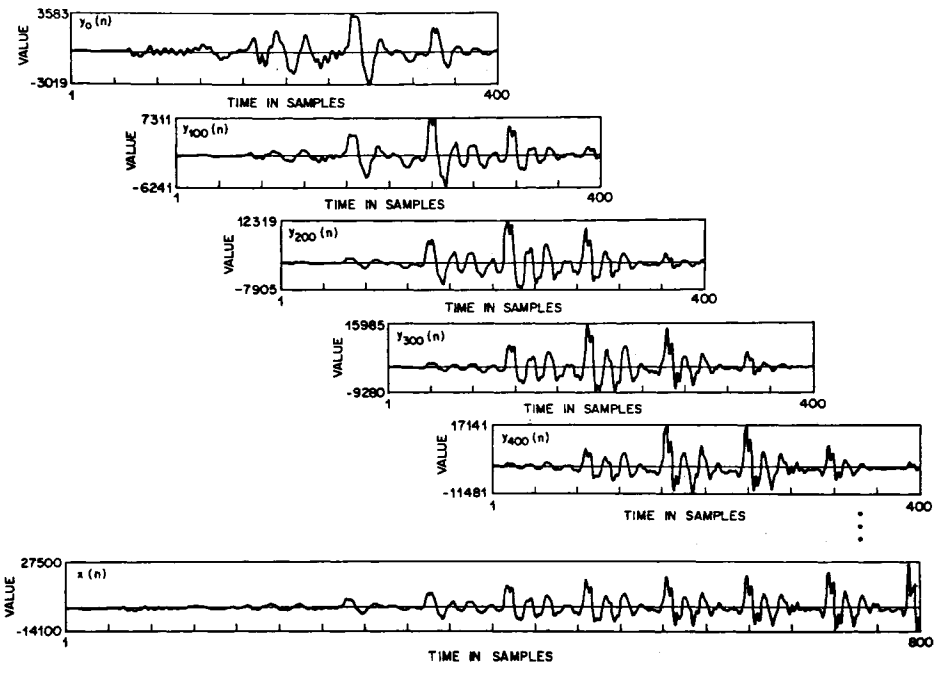

Fig. 3. Graphical interpretation of the overlap-add synthesis method showing the overlapping sections (weighted by a Hamming window) and the resulting summation.

then $x(m)$ (or equivalently $x(n)$) can be reconstructed by taking the inverse Fourier transform of $X_n(e^{j\omega k})$ and dividing out the window (assuming it is nonzero for all values of n which are considered). Although this procedure is valid it will be shown later that the resulting short-time transform estimates are an undersampled (in time) representation of the signal, and as such turn out not to be useful for applications when modifications are made to the short-time transform. In this section we present a properly sampled and more robust synthesis procedure similar to the OLA method [13], [14] (which does aperiodic convolution using discrete Fourier transforms).

The synthesis procedure for the OLA method is to form the signal

$$y(n) = \sum_m \left[\sum_k X_m(e^{j\omega k}) e^{j\omega k n} \right] \tag{4.2}$$

where the summation over m is for overlapping analysis sections with short-time Fourier transform $X_m(e^{j\omega k})$. Basically (4.2) says that to reconstruct the signal we inverse Fourier transform $X_m(e^{j\omega k})$ for each m at which an analysis was performed, which, by the definition of X, gives

$$y_m(n) = Lx(n)\, w(m - n) \tag{4.3}$$

where L is the size of the inverse discrete Fourier transform and then we sum $y_m(n)$ over m giving

$$y(n) = \sum_m y_m(n) = Lx(n) \sum_m w(m - n). \tag{4.4}$$

Fig. 3 illustrates the summation of (4.4) for the case of $w(n)$ a Hamming window of duration N samples, and for short-time analyses being performed every $m = N/4$ samples. It is seen that, given any value of n, a total of 4 distinct $y_m(n)$ terms contribute to the value of $y(n)$ for this example.

The term $\Sigma_m w(m - n)$ of (4.4) is seen to be the sum of the window shifted by m samples. By recognizing that the expression $\Sigma_m\ w(m - n)$ is simply a sum of the values of a low-pass window, it can be shown [8] that if $w(n)$ is sampled at a sufficiently dense rate, then

$$\sum_m w(m - n) = W(e^{j0}) \tag{4.5}$$

independent of the window offset n, where $W(e^{j0})$ is the value of $W(e^{j\omega})$, the transform of the window, evaluated at zero frequency. Thus (4.4) becomes

$$y(n) = Lx(n)\, W(e^{j0}) \tag{4.6}$$

showing that the synthesis rule of (4.2) will lead to exact reconstruction of $x(n)$ by overlap-adding sections of the waveform.

The entire synthesis procedure depends on the sampling relation of (4.5). This relationship is valid to within an aliasing error which can be made negligiably small for sufficiently high sampling rates of the window—i.e., as the sampling rate of the short-time Fourier transform estimates increases, the aliasing error decreases monotonically to zero.

V. Effects of Modifications to the Short-Time Transform on the Resulting Synthesis

At this point we have shown that there are two distinctly different methods for reconstructing a signal from its short-time Fourier transform. Both methods have been shown to be capable of reconstructing the original signal exactly (within a scale factor which is different in each case) when the short-time transform is unmodified. For most (if not all) applications, however, one is interested in making modifications to the short-time Fourier transform. These modifications take on the form of truncation errors in vocoder applications and time-

varying filtering for signal processing applications. In this section we show the effects on the synthesized signal of fixed and time varying multiplicative modifications to the short-time transform.

A. FBS Method

We represent a fixed multiplicative modification to the short-time transform as

$$\hat{Y}_n(e^{j\omega_k}) = X_n(e^{j\omega_k})\, P(e^{j\omega_k}) \tag{5.1.1}$$

where $P(e^{j\omega_k})$ is a frequency weighting function on the short-time transform. We assume that the inverse transform of $P(e^{j\omega_k})$ exists, and we call this sequence $p(n)$ where [2]

$$p(n) = \frac{1}{L} \sum_k P(e^{j\omega_k})\, e^{j\omega_k n} \tag{5.1.2}$$

and L is the number of frequencies at which $P(e^{j\omega_k})$ is evaluated—i.e., the number of analysis frequencies. The reconstructed signal $\hat{y}(n)$ from the FBS method is

$$\hat{y}(n) = \sum_k X_n(e^{j\omega_k})\, P(e^{j\omega_k})\, e^{j\omega_k n} \tag{5.1.3}$$

$$= \sum_k \left[\sum_m w(n-m)\, x(m)\, e^{-j\omega_k m} \right] P(e^{j\omega_k})\, e^{j\omega_k n} \tag{5.1.4}$$

$$= \sum_m w(n-m)\, x(m) \sum_k P(e^{j\omega_k})\, e^{j\omega_k (n-m)} \tag{5.1.5}$$

$$= \sum_m w(n-m)\, x(m)\, Lp(n-m) \tag{5.1.6}$$

$$= Lx(n) * [w(n)\, p(n)]. \tag{5.1.7}$$

Thus the effect of the fixed spectral modification $P(e^{j\omega_k})$ is to convolve the signal $x(n)$ with the product of the window $w(n)$ and the impulse response of the modification $p(n)$. Ideally one would expect the result to be of the form

$$\tilde{y}(n) = x(n) * p(n) \tag{5.1.8}$$

rather than of the form of (5.1.7). Thus for the FBS method, fixed spectral modifications are strongly affected by the window, and only in the case when the time duration of $p(n)$ is short compared to the window duration is it even approximately true that

$$\hat{y}(n) \approx \tilde{y}(n) \tag{5.1.9}$$

For time-varying modifications we model $\hat{Y}_n(e^{j\omega_k})$ as

$$\hat{Y}_n(e^{j\omega_k}) = X_n(e^{j\omega_k})\, P_n(e^{j\omega_k}) \tag{5.1.10}$$

and we define the time-varying impulse response due to the modification $p_n(m)$ as

$$p_n(m) = \frac{1}{L} \sum_k P_n(e^{j\omega_k})\, e^{j\omega_k m}. \tag{5.1.11}$$

Proceeding as before we solve for $\hat{y}(n)$, due to the modification, as

$$\hat{y}(n) = \sum_k X_n(e^{j\omega_k})\, P_n(e^{-j\omega_k})\, e^{j\omega_k n} \tag{5.1.12}$$

$$= \sum_k e^{j\omega_k n} \sum_k x(n-m)\, w(m)\, e^{j\omega_k m} P_n(e^{j\omega_k})\, e^{j\omega_k n} \tag{5.1.13}$$

$$= \sum_m x(n-m)\, w(m) \sum_m P_n(e^{j\omega_k})\, e^{j\omega_k m} \tag{5.1.14}$$

$$= \sum_m x(n-m)\, w(m)\, Lp_n(m) \tag{5.1.15}$$

$$= L \sum_m x(n-m)\, [p_n(m)\, w(m)]. \tag{5.1.16}$$

Equation (5.1.16) shows that for the FBS method the time response of the spectral modification is weighted by the window before being convolved with $x(n)$. Note also that the effect of the spectral modification is instantaneous in time.

In summary, for the FBS method, the effect of a spectral modification (either fixed or time-varying) is to convolve the original signal with a time-limited window-weighted version of the time response due to the modification. As such this synthesis method would be useful for applications in which modifications were being made where the time response due to the modification (i.e., $p_n(m)$) might be *uncontrollably* long. Although the resulting modifications do not match those which were intended in this method, undesired large smearing in time of the signal due to the modification is controlled in the FBS method. Further, the time fidelity of the modification is maintained.

B. OLA Method

Using the representation of (5.1.1) for the OLA modification we obtain for the reconstructed signal

$$\hat{y}(n) = \sum_m \sum_k X_m(e^{j\omega_k})\, P(e^{j\omega_k})\, e^{j\omega_k n} \tag{5.2.1}$$

$$= \sum_m \sum_k \sum_l x(l)\, w(m-l)\, e^{-j\omega_k l} P(e^{j\omega_k})\, e^{j\omega_k n} \tag{5.2.2}$$

$$= \sum_l x(l) \left[\sum_k P(e^{j\omega_k})\, e^{j\omega_k (n-l)} \right] \left[\sum_m w(m-l) \right] \tag{5.2.3}$$

$$= \sum_l x(l)\, Lp(n-l)\, W(e^{j0}) \tag{5.2.4}$$

or

$$\hat{y}(n) = LW(e^{j0})\, [x(n) * p(n)]. \tag{5.2.5}$$

Equation (5.2.5) shows that $\hat{y}(n)$ *is the convolution of the original signal with the time response of the spectral modification*—i.e., no window modifications of $p(n)$ have occurred with OLA. (The reader should realize that an appropriate change must be made to the analysis—i.e., padding the windowed input signal with a sufficient number of zero valued samples—to prevent time aliasing when implementing the analysis and synthesis operations with FFT's, which have length L. If a modification $P(e^{j\omega_k})$ has a time response which is effectively N_0 points long, the analysis length L must be at least $N + N_0 - 1$ where the window length is N.)

For the case of a time-varying modification we obtain

$$\hat{y}(n) = \sum_m \left[\sum_k X_m(e^{j\omega_k})\, P_m(e^{j\omega_k}) \right] e^{j\omega_k n} \tag{5.2.6}$$

[2] It is assumed that for all inverse DFT's the sequence is 0 outside the range of the DFT index.

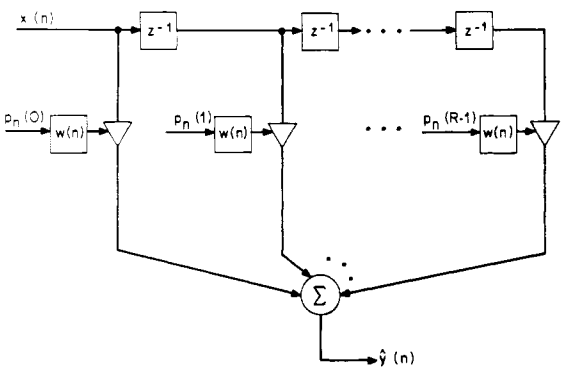

Fig. 4. Block diagram of the effects (in time) of making modifications to the short-time spectrum for the OLA method.

which can be manipulated into the form

$$\hat{y}(n) = \sum_l x(l) \sum_m w(m-l) \left[\sum_k P_m(e^{j\omega_k}) e^{j\omega_k(n-l)} \right]. \tag{5.2.7}$$

Using (5.1.11), we get

$$\hat{y}(n) = L \sum_l x(l) \left[\sum_m w(m-l)\, p_m(n-l) \right]. \tag{5.2.8}$$

If we let $r = n - l$ or $l = n - r$ then (5.2.7) becomes

$$\hat{y}(n) = L \sum_r x(n-r) \sum_m p_m(r)\, w(m-n+r). \tag{5.2.9}$$

If we define $\hat{p}$ by

$$\hat{p}(r-n, r) = \hat{p}(q, r) = \sum_m p_m(r)\, w(m-q) \tag{5.2.10}$$

then (5.2.8) becomes

$$\hat{y}(n) = L \sum_r x(n-r)\, \hat{p}(r-n, r). \tag{5.2.11}$$

The interpretation of (5.2.10) is that for r held constant $\hat{p}(q, r)$ is the true convolution of $p_m(r)$ and $w(m)$. When (5.2.11) is interpreted in the time domain, OLA is equivalent to a tapped-delay line with time-varying tap weights where each tap weight is bandlimited by the low-pass window w. Fig. 4 shows a simple interpretation of this result.

Thus for the overlap add method, the time varying responses of each tap due to the spectral modifications are *bandlimited by the window* but the modification acts as a true convolution on the input signal. This is in direct contrast to the FBS method in which the modifications were *time limited* by the window and could change instantaneously.

VI. Additive Modifications

We have been discussing the effects of nonrandom multiplicative modifications to the short-time transform. It is also important to understand the effects of additive random modifications to the short-time transform. This type of modification will occur when implementing the analysis with finite precision (i.e., roundoff noise), or when quantizing the short-time transform as for a vocoder [4].

We model additive modifications to the short-time Fourier transform as

$$\hat{Y}_n(e^{j\omega_k}) = X_n(e^{j\omega_k}) + E(e^{j\omega_k}) \tag{6.1}$$

where we define the inverse Fourier transform of $E(e^{j\omega_k})$ as

$$e(n) = \frac{1}{L} \sum_k E(e^{j\omega_k})\, e^{j\omega_k n}. \tag{6.2}$$

(In the case where $e(n)$ is a random noise, then a statistical model for $e(n)$ and $E(e^{j\omega_k})$ might be used. The results to be presented are not dependent on such a statistical model.)

For the FBS method the effect of the additive modification of (6.1) is

$$\hat{y}(n) = \sum_k (X_n(e^{j\omega_k}) + E(e^{j\omega_k}))\, e^{j\omega_k n} \tag{6.3}$$

which, by linearity, can be put in the form

$$\hat{y}(n) = y(n) + \sum_k E(e^{j\omega_k})\, e^{j\omega_k n} \tag{6.4}$$

or

$$\hat{y}(n) = y(n) + Le(n) \tag{6.5}$$

where $y(n)$ is as defined in (3.8). Thus an additive spectral noise modification results in an additive noise component in the reconstructed signal. The reader should notice that the analysis window has no direct effect on the additive term in the synthesis but that the noise increases linearly with the number of bands L.

For the OLA method the effect of the additive modification of (6.2) is

$$\hat{y}(n) = \sum_m \sum_k (X_m(e^{j\omega_k}) + E(e^{j\omega_k}))\, e^{j\omega_k n} \tag{6.6}$$

which can be put in the form

$$\begin{aligned} \hat{y}(n) &= y(n) + \sum_m \left[\sum_k E(e^{j\omega_k})\, e^{j\omega_k n} \right] \\ &= y(n) + L \sum_m e(n) \end{aligned} \tag{6.7}$$

where $y(n)$ is defined in (4.6). Thus for additive modifications the resulting synthesis contains a larger additive signal for the OLA method than for the FBS method due to the overlap between analysis frames. For a Hamming window with a 4-to-1 overlap where L the FFT length equals the window length, the additive term in the synthesis will be on the order of four times greater[3] (twice the noise power) for the OLA method than for the FBS method. As such the OLA method tends to introduce more noise than the FBS method, and thus would be less useful for vocoding applications, etc. It is easy to understand these results by recognizing that the noise component $e(n)$ will always time alias regardless of the FFT length. This violates a basic assumption of the OLA synthesis method, namely that the short-time Fourier transform was adequately sampled in frequency.

VII. Sampling Rates of $X_n(e^{j\omega_k})$ in Time and Frequency

A basic consideration in the implementation of systems for short-time Fourier analysis and synthesis is the selection of the rate at which $X_n(e^{j\omega_k})$ should be sampled in both time (n)

[3] We are assuming that noise sequences for consecutive frames are uncorrelated. This is approximately true for most practical implementations.

and frequency (k) to provide an unaliased representation of $X_n(e^{j\omega_k})$. This question requires a careful consideration of the factors entering into the computation of $X_n(e^{j\omega_k})$. Unfortunately confusion has existed in the past on this point which has masked the real issues. The confusion is a result of the fact that if sampling rates lower than the theoretically minimum rate are used in either time (for the OLA method), or frequency (for the FBS method), $x(n)$ can still be *exactly recovered* from the aliased (undersampled) short-time transform when no modifications have been made. Such undersampled representations are actually quite useful for applications in which one is only interested in obtaining the short-time transform (e.g., spectral estimation, parameter estimation, etc.), for vocoder applications in which minimization of overall bit rate of the system is of prime importance, and for convolution by FFT methods. However, for applications in which one is interested in obtaining a short-time Fourier transform of the signal, performing some modification of the spectrum (e.g., fixed or time-varying filtering), and then synthesizing the modified signal, it is essential that little or no aliasing occur in either the time or frequency domains.

First we will discuss the required sampling rate of $X_n(e^{j\omega_k})$ in time. In this case, the linear filtering interpretation of Section II provides the necessary insight. There it was shown that for a fixed value of ω_k, $X_n(e^{j\omega_k})$ was the output of a filter with impulse response $w(n)$. We have assumed from the beginning that $W(e^{j\omega})$, the Fourier transform of $w(n)$, is a lowpass function of bandwidth B Hz. Therefore, the frequency bandwidth of $X_n(e^{j\omega_k})$ is the same as that of the window, and thus according to the sampling theorem, $X_n(e^{j\omega_k})$ must be sampled at a rate of at least $2B$ samples per second (sampling period of $1/(2B)$ second) to avoid aliasing. By way of example, for $w(n)$ a Hamming window of length N samples, then the bandwidth B is

$$B = \frac{2F_s}{N} \text{ (Hz)} \tag{7.1}$$

where F_s is the sampling rate of the signal $x(n)$. Therefore, the required sampling rate of $X_n(e^{j\omega_k})$ in time is $2B = 4F_s/N$ samples per second. Thus for $N = 100$, $F_s = 10\,000$ Hz, we get $B = 200$ Hz, and we require $X_n(e^{j\omega_k})$ to be evaluated 400 times per second—i.e., every 25 samples. In general the sampling rate for an N-point Hamming window is $N/4$, based on a 42-dB criterion on the log magnitude spectrum—i.e., the bandwidth B is defined as the lowest frequency for which the log magnitude spectrum remains at least 42 dB below the peak value.

For the OLA method, we have already shown that the analysis need be performed only $2B$ times per second, and that the synthesis method reconstructs the signal by overlap-adding the individual time responses due to each analysis frame (with the appropriate time shift). For the FBS synthesis method we require $X_n(e^{j\omega_k})$ to be known for each sample at the sampling rate of the original signal F_s. As such, interpolation methods must be used to interpolate $X_n(e^{j\omega_k})$ from a rate of $2B$ samples per second to the rate of F_s samples per second as required by the synthesis procedure. Methods for performing the interpolation are discussed in [4] and [7].

To determine the required sampling rate of $X_n(e^{j\omega_k})$ in frequency, i.e., to determine a finite set of frequencies $\omega_k = 2\pi k/L$, $k = 0, 1, \cdots, L-1$, at which $X_n(e^{j\omega_k})$ must be specified to exactly recover $x(n)$, we use the Fourier transform interpretation of $X_n(e^{j\omega_k})$. Since the inverse Fourier transform of $X_n(e^{j\omega_k})$ is time-limited, we can use the sampling theorem to sample $X_n(e^{j\omega_k})$ in frequency at a rate of at least twice this "time width."[4] Since the inverse Fourier transform of $X_n(e^{j\omega_k})$ is the signal $x(m)\,w(n-m)$, and this signal is of duration N samples (due to the finite duration window $w(n)$), then the sampling theorem says that $X_n(e^{j\omega_k})$ must be sampled at the set of frequencies corresponding to the N roots of unity

$$\omega_k = \frac{2\pi k}{N}, \quad k = 0, 1, \cdots, (N-1) \tag{7.2}$$

in order to exactly recover $x(n)$ from $X_n(e^{j\omega_k})$. Thus for our example of a Hamming window of duration $N = 100$ samples, we require $X_n(e^{j\omega_k})$ to be evaluated at 100 uniformly spaced frequencies around the unit circle. When modifications have extended the time length, ω_k must be appropriately increased in the OLA case. (In FBS, modifications cannot increase the time length.)

Based on the above discussion the *total* number of samples of $X_n(e^{j\omega_k})$ that must be computed per second for $w(n)$, an N-point Hamming window, is

$$NT = N\,\frac{4F_s}{N} = 4F_s \tag{7.3}$$

or the ratio between NT and F_s is

$$R = \frac{NT}{F_s} = 4. \tag{7.4}$$

Thus for our example of a Hamming window, a properly sampled short-time transform requires on the order of 4 times more information as would be required relative to the original signal $x(n)$. When modifications are to be made, this represents a lower bound on the required information rate for the short-time transform. In return for this redundancy one obtains a very flexible signal representation for which extensive modifications in both the time and frequency dimensions can be made.

We have already discussed the case in which $X_n(e^{j\omega_k})$ can be undersampled in time for the OLA method (Section IV). Similarly $X_n(e^{j\omega_k})$ can be undersampled in frequency for the FBS method. Basically, for this case, one must design a window whose frequency response approximates an ideal lowpass filter as closely as possible. Then the number of frequency bands can be reduced to the minimum by using contiguous (nonoverlapping) analysis bands. A reduction in the number of analysis bands of 4 to 1 as compared to a Hamming window analysis can be obtained in this manner. Details of this implementation are discussed in [4]–[7].

VIII. Duality Between FBS and OLA

Throughout this paper we have illustrated the complementary nature of the two synthesis methods. We now show that, as a result of properties of the window $w(n)$, a formal duality exists. The duality is based on the simple relations

$$\sum_m w(n-m) = W(e^{j0}) \quad \text{any } n \tag{8.1}$$

[4] The definition of the "time width" of the window is the *total* duration of the window, whereas the "bandwidth" of the window is defined as the cutoff frequency of the window.

$$\frac{1}{L}\sum_{k} W(e^{j(\omega-\omega_k)}) = w(0) \qquad \text{any } \omega \qquad (8.2)$$

where L is the number of values of k used in the summation. Equation (8.1) says that if $w(n)$ is adequately sampled in time (i.e., for sufficiently many values of m) then the sum of the sampled values of $w(n)$ is the 0 frequency value of the Fourier transform of the window $W(e^{j0})$ independent of the time shift n. Conversely, Equation (8.2) says that if $W(e^{j\omega})$ is adequately sampled in frequency then the sum of the sampled values of $W(e^{j\omega})$ is the zero time value of the inverse Fourier transform of the frequency response of the window, i.e., $w(0)$.

We have already shown that (8.1) forms the basis for the OLA synthesis method (4.5). It is readily seen that (8.2) forms the basis for the FBS synthesis method by taking the Fourier transform of (3.9), and recalling that the multiplication of a sequence by $e^{j\omega_k n}$ in time corresponds to shifting the Fourier transform of the sequence by ω_k in frequency.

Based on the duality relations of (8.1) and (8.2), it is easy to show that either synthesis method can be derived from the other method by replacing each dependent variable by its Fourier transform, and then interchanging the roles of frequency and time.

Finally it should be noted that (8.1) and (8.2) may be derived from the Poisson sum formula [15] by using either the "time limited" or bandlimited properties of the window [8].

IX. Nonlinear Modifications

The idea of applying nonlinear modifications of the short-time Fourier transform prior to synthesis is a new unexplored area in signal processing. Several important applications are known to exist at the present time. One is the speech stretching problem where the time structure of a speech signal is stretched without modifying the pitch. This application is one in which results from the OLA method appear to be better than with other methods. For example, to stretch speech by a factor of two, a long window is used (25.6 ms) in order to resolve each pitch harmonic. Then the number of short-time transforms are doubled by linear (or bandlimited) interpolation giving new short-time transforms at twice the time sampling rate of the original short-time transform. Finally the phase is computed at each frequency and doubled (the nonlinear part of the calculation). The stretched speech is then synthesized by the OLA method. The analysis-synthesis procedure is done using an FFT with twice the length of the window to allow for the time response due to the modifications which have been made. Informal comparisons of the synthesis from the OLA method with that of the FBS method show the OLA method is better suited to this application.

Another application for nonlinear spectral modifications is the speech dereverberation problem in which time varying modifications to the short-time spectrum are dependent on the short-time spectrum itself. Details of this application are given in [3]. Several other applications of nonlinear modifications to the short-time Fourier transform including dynamic range compression of speech and noise removal in music are described by Callahan [12].

From experience it is known that the OLA method works well when the modifications being made are a function of the short-time transform. This is viewed as a nonlinear modification since the coefficients of the time-varying filter are derived from the signal. The power of the technique is that linear system ideas may be meaningfully applied to a nonstationary, nonlinear problem.

X. Summary

The purpose of this paper was to unify the various approaches to implementing systems for short-time Fourier analysis and synthesis of a signal. We have discussed the similarities and differences between the two proposed methods of synthesizing a signal from its short-time transform. Finally, it was shown that a formal duality between the two synthesis methods could be stated which clearly displays the complementary nature of the two techniques.

Acknowledgment

The authors wish to acknowledge the contributions of Dr. J. L. Flanagan and Professor R. W. Schafer to the initial development of the FBS method, to Dr. M. R. Portnoff and Dr. D. A. Berkley for several useful discussions on the complementary nature of the two synthesis methods.

References

[1] J. L. Flanagan and R. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493-1509, 1966.
[2] J. L. Flanagan and R. C. Lummis, "Signal processing to reduce distortion in small rooms," *J. Acoust. Soc. Amer.*, vol. 47, no. 6, pp. 1465-1481, June 1970.
[3] J. B. Allen, D. A. Berkley, and J. Blauert, "A multi-microphone signal processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 61, Oct. 1977.
[4] R. W. Schafer and L. R. Rabiner, "Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 3, pp. 165-174, June 1973.
[5] ——, "Design of digital filter banks for speech analysis," *Bell Syst. Tech. J.*, vol. 50, no. 10, pp. 3097-3115, Dec. 1971.
[6] R. W. Schafer, L. R. Rabiner, and O. Herrmann, "FIR digital filter banks for speech analysis," *Bell Syst. Tech. J.*, vol. 54, no. 3, pp. 531-544, Mar. 1975.
[7] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. ASSP-24, no. 3, pp. 243-248, June 1976.
[8] J. B. Allen, "Short-term spectral analysis and synthesis and modification by discrete Fourier transform," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. ASSP-25, no. 3, pp. 235-238, June 1977.
[9] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, pp. 57-62, Aug. 1970.
[10] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental frequency measurements," *J. Acoust. Soc. Amer.*, vol. 43, no. 4, pp. 829-834, Apr. 1968.
[11] M. V. Mathews, J. E. Miller, and E. E. David, Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, pp. 179-186, 1961.
[12] M. J. Callahan, "Acoustic signal processing based on the short-time spectrum," Univ. Utah Rep. CSc-76-209, Mar. 1976.
[13] T. G. Stockham, "High speech convolution and correlation," in *Proc. AFIPS Spring Joint Computer Conf.*, vol. 28, pp. 229-233, 1977.
[14] H. D. Helms, "Fast Fourier transform method of computing difference equations and simulating filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 85-90, June 1967.
[15] A. Papoulis, *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962, pp. 47-50.