HST.582J / 6.555J / 16.456J Biomedical Signal and Image Processing
Spring 2007

Chapter 7 - TIME-DEPENDENT PROCESSING OF SIGNALS THE SHORT-TIME FOURIER TRANSFORM

©Bertrand Delgutte 1999

Introduction

In the preceding chapters, we have introduced methods for analyzing two broad classes of signals. The first class is that of *finite-energy* signals. Such signals rapidly decay to zero outside of a finite time interval, so that they have well-defined Fourier transforms. The second class of signals is that of *stationary,* random signals, which last indefinitely, and have stable statistical characteristics over long times. Examples of stationary signals are periodic signals and "white noise", the output of a random number generator. Because such signals are of infinite duration, their Fourier transforms are, in general, not mathematically defined. However, they have a *power spectrum* obtained by Fourier transforming the autocorrelation function. Because the autocorrelation function is a time average, the power spectrum represents the average frequency content of the signal over all times. There is a third type of signals for which neither Fourier transforms nor power spectra are applicable. These *nonstationary signals* are of indefinite duration, so that their Fourier transforms may not exist, and have statistical characteristics that change appreciably over time, so that it would not make sense to compute an average power spectrum for all times. For example, speech signals have spectral characteristics that change continuously over time. These variations are key for communication because a stationary signal cannot convey new information. Electrocardiographic signals also show temporal variations in statistical characteristics that reflect changes in the state of the heart. These changes are important in the clinic, for example in the diagnosis of arryhtmias.

In this chapter, we introduce techniques for processing a special class of nonstationary signals. The basic idea is to divide the signal into short time segments or "frames" over which the signal is approximately stationary, then make a set of measurements for each frame. Such *time-dependent processing* is applicable whenever the signal is *quasi-stationary,* i.e. when its statistical characteristics change slowly relative to the frame length. For example, speech can be considered to be quasi-stationary because the motions of the the articulators are usually sufficiently slow that the spectral characteristics of speech change relatively little over intervals of 10 to 30 ms.

The most important type of time-dependent processing is *short-time Fourier analysis,* which gives the energy distribution of a signal as a function of both time and frequency. Short-time Fourier analysis of speech signals is used for tracking important parameters such as the formant frequencies. It is also the basis for generating spectrographic displays and for vocoder systems that provide efficient storage and transmission of speech and audio signals. By varying the duration of the frames used for short-time Fourier analysis, frequency resolution can be traded for time resolution in the resulting energy distribution.

7.1 Time-dependent processing

7.1.1 Definition

A time-dependent measurement T_n on a signal $x[n]$ is obtained by first applying a transformation $T(.)$ (which is usually nonlinear) to the signal, and then lowpass filtering the transformed signal:

$$T_n \triangleq \sum_{m=-\infty}^{\infty} T(\ x[m]\ )\ w[n-m]\ =\ T(\ x[n]\ )\ *\ w[n] \tag{7.1}$$

Figure 1a shows a block diagram representation of time-dependent processing. Time-dependent measurements depend not only on the transformation $T(.)$, but also on the choice of the window function $w[n]$. This window function is the unit-sample response of the lowpass filter. In most applications, the window has a finite duration, so that the infinite sum in (7.1) is in fact a local weighted average of $T(\{x[m]\})$ centered at time n. The output T_n varies with time, but more slowly than does the signal $x[n]$ because $w[n]$ is a lowpass filter. Thus, it is usually not necessary to evaluate T_n for every time sample n: By the Nyquist theorem, it suffices to evaluate T_n at intervals of $1/2\Delta W$ samples, where ΔW is the cutoff frequency of the lowpass filter $w[n]$. Effectively, the bandwidth of the window defined the interval between frames used in short-time analysis.

7.1.2 Short-time energy

A useful example of (7.1) is obtained when the transformation $T(.)$ is a square function: The output signal is then the *short-time energy* E_n of the signal $x[n]$:

$$E_n \triangleq \sum_{m=-\infty}^{\infty} x[m]^2\ w_0[n-m] \tag{7.2}$$

The window $w_0[n]$ must be positive in order to guarantee that the short-time energy be positive for all times. In the special case when $w_0[n]$ is a rectangular pulse of length $2N+1$ centered at the origin, this becomes

$$E_n\ =\ \sum_{m=n-N}^{n+N} x[m]^2$$

which corresponds to the intuitive notion of signal energy over the interval $[n-N,\ n+N]$. Figure 2 shows the short-time energy of a speech signal for Hamming windows of different lengths. Long windows provide a smooth short-time energy, while short windows are best for resolving fine temporal variations in signal energy. Sometimes one is interested in the short-time energy *over a specific frequency band* rather than that of the entire signal. In this case, the signal is first bandpass filtered, then the short-time energy of the filter output is computed.

Direct application of the definition (7.2) gives one possible method for computing the short-time energy: The input signal is squared, then processed by a lowpass filter with unit-sample response $w_0[n]$ (Fig. 1b). This method is the best one when $w_0[n]$ is an IIR filter because the short-time energy can be evaluated recursively. On the other hand, if $w_0[n]$ is of finite duration, there exists an alternative method for computing the short-time energy that is often more efficient. This alternative method can be derived by defining a new window $w[n] \overset{\triangle}{=} \sqrt{w_0[-n]}$, then making the change of variable $l = m - n$ in (7.2):

$$E_n = \sum_{m=-\infty}^{\infty} x[m]^2 w[m-n]^2 = \sum_{l=-\infty}^{\infty} (x[n+l]w[l])^2 \tag{7.3}$$

A block diagram representation of this method is shown in Fig. 1c. For each time n, this method amounts to computing the total energy in the windowed signal $x_n[m] \overset{\triangle}{=} x[n+m]\ w[m]$. This method is particularly efficient when the short-time energy needs only be computed for certain times, e.g. at multiples of $1/2W$.

7.1.3 Short-time autocorrelation function

A generalization of the short-time energy is the *short-time autocorrelation function* $R_n[k]$, which is a function of two variables, time n and lag k. It is obtained by premultiplying the portion of the signal centered at time n by a window function, forming $x_n[m] \overset{\triangle}{=} x[n+m]\ w[m]$, then computing the deterministic (a.k.a. "raw") autocorrelation function of the windowed signal $x_n[m]$ (Fig. 1e):

$$R_n[k] \overset{\triangle}{=} x_n[k] * x_n[-k] \;=\; \sum_{m=-\infty}^{\infty} x_n[m]\, x_n[m-k] \;=\; \sum_{m=-\infty}^{\infty} x[n+m]\, w[m]\, x[n+m-k]\, w[m-k] \tag{7.4}$$

For example, if $w[n]$ is a rectangular pulse of length $2N+1$ centered at the origin, one has:

$$R_n[k] \;=\; \sum_{m=-(N-|k|)}^{N} x[n+m]\ x[n+m-|k|]$$

Note that the summation is over $2N - |k| + 1$ terms, so that the short-time autocorrelation function is zero for $|k| > 2N$. This is true for all windows of length $2N+1$.

Comparing (7.4) with (7.3) shows that $R_n[0] = E_n$ if the same window $w[n]$ is used in both expressions. In fact, because $R_n[k]$ is the deterministic autocorrelation function of the windowed signal $x_n[m]$, it has all the properties of autocorrelation functions:

$$R_n[k] \;=\; R_n[-k]$$

$$|R_n[k]| \;\leq\; R_n[0] \;=\; E_n$$

If the signal is periodic over the duration of the window, the short-time autocorrelation function has local maxima when the lag k is a multiple of the period. Thus, the short-time autocorrelation function is useful for tracking the period of quasi-periodic signals (such as voiced speech) whose period slowly changes over time (Fig. 3).

The short-time autocorrelation function can be written in the form of time-dependent processing (7.1) if we define a new filter $w_k[n]$ analogous to $w_0[n]$ in (7.2). For this purpose, we make the change of variable $l = n + m$ in (7.4):

$$R_n[k] \;=\; \sum_{m=-\infty}^{\infty} x[n+m]\, x[n+m-k]\, w[m]\, w[m-k] \;=\; \sum_{l=-\infty}^{\infty} x[l]\, x[l-k]\ [w[l-n]\ w[l-n-k]]$$

If, for each lag k, we introduce the new window function

$$w_k[n] \ \overset{\triangle}{=} w[-n] \ w[-k-n] \tag{7.5a}$$

the short-time autocorrelation function becomes

$$R_n[k] \ = \ \sum_{l=-\infty}^{\infty} \ (x[l] \ x[l-k]) \ w_k[n-l] \ = \ (x[n] \ x[n-k]) \ * \ w[n] \tag{7.5b}$$

This expression is in the form (7.1), with the transformation $T(.)$ being multiplication of the signal $x[n]$ by the delayed signal $x[n-k]$. This alternative interpretation of the short-time autocorrelation function is illustrated in Fig. 1d. Note that the new window $w_k[n]$ is different for each lag, and that its length is $N - |k|$ if $w[n]$ is of length N.

There are many other forms of time-dependent processing besides the short-time energy and the short-time autocorrelation function. In fact, any time-average operation on stationary signals can be extended to a form of time-dependent processing for nonstationary signals by substituting a lowpass filter for the infinite time average. The most important form of time-dependent processing is the short-time Fourier transform, which can be seen as a generalization of the power spectrum to nonstationary signals.

7.2 The short-time Fourier transform

7.2.1 Definition

The *short-time Fourier transform* $X_n(f)$ is function of two variables, time n and frequency f, which describes how the spectrum of restricted segments of a signal $x[n]$ evolves with time. Formally, it is defined by:

$$X_n(f) \ \overset{\triangle}{=} \sum_{m=-\infty}^{\infty} \ x[m] \ w[n-m] \ e^{-j2\pi fm} \tag{7.6}$$

While this definition may appear somewhat daunting, it can be interpreted from at least three different points of view (Fig. 4): a *time-dependent processing* interpretation, a *Fourier transform* interpretation, and a *filter-bank* interpretation. Each of these interpretations provides different insights into the properties of short-time Fourier analysis, and leads to alternative implementations that can be particularly efficient in certain applications.

7.2.2 Time-dependent processing interpretation

The short-time Fourier transform (STFT) is a form of time-dependent processing (7.1) in which the transformation $T(.)$ is multiplication of the signal by a complex exponential. To show this, we start from the definition (7.6), and hold the frequency f at a specific value f_0. When considered as a function of n, $X_n(f_0)$ can be written in the form of a convolution:

$$X_n(f_0) = \Big[\ x[n] \ e^{-j2\pi f_0 n} \ \Big] \ * \ w[n] \tag{7.7}$$

This represents the cascade of a modulation (multiplication) by the complex exponential $e^{-j2\pi f_0 n}$ and lowpass filtering by the window $w[n]$ (Fig. 7.4a). From the product theorem, modulation translates the signal spectrum $X(f)$ (assumed to exist) by $-f_0$, bringing the frequency components near f_0 down to DC:

$$x[n]e^{-j2\pi f_0 n} \longleftrightarrow X(f+f_0) = X(f) \circledast \tilde{\delta}(f+f_0)$$

The convolution of the modulated signal by the lowpass window $w[n]$ then retains these frequency components in the output, and rejects all other components. From these observations, we deduce two key properties of $X_n(f_0)$ considered as function of time n:

1. It is a lowpass signal, in the sense that it varies with n much slower than does the original signal $x[n]$. The bandwidth of $X_n(f_0)$ is the same as that of $w[n]$.
2. It contains the frequency components of $x[n]$ that are within $\pm\Delta W$ of f_0, where ΔW is the bandwidth of $w[n]$.

The time-dependent processing interpretation of the STFT is important because it places the STFT within the general framework of this chapter. However, unlike the other two interpretations, it is rarely used as a basis for implementations of the STFT.

7.2.3 Fourier transform interpretation

The Fourier transform interpretation of the STFT is the basis for modern implementations of the STFT. It is most easily understood by considering $X_n(f)$ as a function of frequency for a specific time $n = n_0$. From the definition (7.6), it is clear that $X_{n_0}(f)$ is the discrete-time Fourier transform of the windowed signal $x_{n_0}[m] = x[m]\ w[n_0 - m]$, as shown in Fig. 7.4b. Therefore, by the product theorem, the STFT is the cyclic convolution of the transform of the signal $x[m]$ (assumed to exist) by the transform of $w[n_0 - m]$, which is $W(-f)\ e^{-j2\pi f n_0}$:

$$x_{n_0}[m] \ = \ x[m]\ w[n_0 - m] \ \longleftrightarrow \ X_{n_0}(f) = X(f) \circledast \left[W(-f)\ e^{-j2\pi f n_0}\right] \qquad (7.8)$$

These operations are shown in Fig. 5 and 6 for a periodic speech-like signal, using Hamming windows of 10 ms and 40 ms respectively. The effect of the window is to "smear" the signal spectrum $X(f)$, so that the frequency resolution of the STFT is limited by the bandwidth of the window ΔW. Specifically, the longer the window, the smaller its bandwidth, and the finer the frequency resolution. In the examples of Fig. 5 and 6, the harmonics of the 100 Hz fundamental frequency are resolved with the 40-ms window, but not with the 10-ms window. Figure 7 shows similar results for an actual speech utterance analyzed with Hamming windows of 5 ms and 50 ms.

The Fourier transform interpretation of the STFT is useful for showing that the signal $x[n]$ can be exactly reconstructed from its STFT. Because $X_n(f)$ is the Fourier transform of the windowed signal $x[m]\ w[n-m]$, the inverse DTFT formula gives

$$x[m]\ w[n-m] \ = \ \int_{-\frac{1}{2}}^{\frac{1}{2}} X_n(f)\ e^{j2\pi f m}\ df$$

Evaluating this expression for $m = n$ yields an exact reconstruction formula for time n (assuming that $w[0] \neq 0$):

$$x[n] = \frac{1}{w[0]} \int_{-\frac{1}{2}}^{\frac{1}{2}} X_n(f) e^{j2\pi fn} df \tag{7.9}$$

This reconstruction formula is not directly useful in practice because it involves an integral over frequency. However, we show in the Appendix that this integral can be replaced by a finite sum (an inverse DFT) under mild conditions on the window $w[n]$.

As a final remark, because the short-time autocorrelation function $R_n[k]$ is the deterministic autocorrelation function of the windowed signal $x_n[k]$, and the STFT is the DTFT of the same windowed signal, the Fourier transform of $R_n[k]$ is the magnitude square of $X_n(f)$:

$$R_n[k] \;=\; x_n[k] \;*\; x_n[-k] \;\longleftrightarrow\; |X_n(f)|^2 \;=\; X_n(f)\; X_n(-f) \tag{7.10a}$$

Applying the inverse DTFT formula to (7.10a) gives:

$$R_n[k] \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} |X_n(f)|^2 \; e^{j2\pi fk} \; df \tag{7.10b}$$

Thus, the magnitude of the STFT (which is often displayed in the form of a spectrogram) contains the same information as the short-time autocorrelation function. Evaluating (7.10b) for $k = 0$ yields Parseval's theorem for the STFT

$$E_n \;=\; R_n[0] \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} |X_n(f)|^2 \; df \tag{7.10c}$$

which simply states that the energy of the windowed signal is conserved in the frequency domain.

7.2.4 Filter-bank interpretation

Historically, the filter-bank interpretation of the STFT was the basis for the first implementations of the STFT in the 1940's. It still plays an important role in modern practice because it is the only practical implementation when the STFT is to be evaluated for arbitrary frequency samples. This interpretation is best understood by holding the frequency f at f_0, and considering $X_n(f_0)$ as a function of time n. To derive the filter-bank formula, we first make the change of variable $l = n - m$ in (7.6):

$$X_n(f_0) \;=\; \sum_{l=-\infty}^{\infty} x[n-l]\; w[l]\; e^{-j2\pi f_0(n-l)} \;=\; e^{-j2\pi f_0 n} \sum_{l=-\infty}^{\infty} x[n-l]\; w[l]\; e^{j2\pi f_0 l} \tag{7.11}$$

It is useful to introduce the *modified short-time Fourier transform:*

$$\tilde{X}_n(f_0) \;\overset{\triangle}{=} X_n(f_0)\; e^{j2\pi f_0 n} \;=\; x[n] \;*\; \Big[\; w[n]\; e^{j2\pi f_0 n} \;\Big] \tag{7.12}$$

Clearly, the STFT and the modified STFT provide the same information, differing only by the multiplicative factor $e^{j2\pi f_0 n}$. In fact, both have the same magnitude: $|\tilde{X}_n(f_0)| = |X_n(f_0)|$. From (7.12), it is apparent that the modified STFT is the convolution of the signal $x[n]$ by the modulated window $w_{f_0}[n] \overset{\triangle}{=} w[n]\ e^{j2\pi f_0 n}$: From the product theorem, modulating the window has the effect of translating its transform $W(f)$ up by f_0:

$$w_{f_0}[n] \ = \ w[n]\ e^{j2\pi f_0 n} \ \longleftrightarrow \ W(f-f_0) \ = \ W(f)\ c*\ \tilde{\delta}(f-f_0)$$

Because the window $w[n]$ has a lowpass spectrum, the transform of the modulated window $W(f-f_0)$ has significant energy only in the vicinity of $f \ = \ f_0$, which means that convolution by the modulated window is a bandpass filtering operation around the center frequency f_0. Figure 8 shows how the DTFT of $\tilde{X}_n(f_0)$ (considered as a function of n) is the product of $X(f)$ (assumed to exist) with the transform of the modulated window $W(f-f_0)$. From these observations, we deduce two properties of the modified STFT $\tilde{X}_n(f_0)$:

1. It is a bandpass signal centered at $f \ = \ f_0$.
2. It consists of the components of the signal $x[n]$ that are within $\pm\Delta W$ of f_0.

From (7.12), the STFT $X_n(f_0)$ can be obtained from the modified STFT $\tilde{X}_n(f_0)$ by demodulation:

$$X_n(f_0) \ = \ \tilde{X}_n(f_0)\ e^{-j2\pi f_0 n}$$

This has the effect of translating the components near f_0 down to DC, thereby creating the lowpass signal $X_n(f_0)$. The effect of this demodulation on the DTFT of $X_n(f_0)$ is shown in Fig. 8. The implementation of the STFT as a cascade of a bandpass filtering operation and a demodulation is shown in Fig. 4c. In order to evaluate the modified STFT for a set of N discrete frequencies f_k, $0 \ \leq \ k \ \leq \ N-1$, the signal $x[n]$ is passed through a bank of N bandpass filters with center frequencies f_k.

The filter-bank interpretation is useful for determining the time resolution of the STFT. Specifically, the modified STFT $\tilde{X}_n(f_0)$ is the convolution of the signal $x[n]$ by the modulated window $w_{f_0}[n]$. This bandpass filter will resolve a pair of pulses in the input providing that its duration (which is the same as that of the original window $w[n]$) is shorter than the separation between the pulses. Thus, the STFT has a time resolution determined by the duration of the window, and a frequency resolution determined by the bandwidth of the window. The uncertainty principle implies that fine time resolution and fine frequency resolution cannot be achieved simultaneously in short-time Fourier analysis.

7.2.5 Efficient implementation of the STFT using the FFT

In practice, the short-time Fourier transform can only be evaluated for a finite number of frequencies. FFT algorithms provide efficient implementations of the STFT when the frequency samples f_k are of the form k/N, where N is a power of two. Specifically, the implementation would be as follows:

1. Multiply the portion of the signal centered at time n by the window.
2. Zero-pad the windowed signal to length N, and compute the N-point DFT of the windowed signal using an FFT algorithm. This gives $X_n(k/N)$ for $0 \ \leq \ k \ \leq \ N-1$.

3. Move the window by R samples, where R is a factor to be determined, and iterate until the end of the signal.

The only unspecified parameter of this method is the sampling interval R, i.e. how frequently the STFT is evaluated in time. We have seen that the STFT is a lowpass signal whose bandwidth is determined by that of the window $w[n]$. According to the Nyquist theorem, if the bandwidth of the window is ΔW, the STFT must be sampled at intervals of $R = 1/(2\Delta W)$ to avoid aliasing. By the uncertainty principle, the bandwidth of the window is inversely related to its duration. For example, the bandwidth of a Hamming window of length N is $2/N$. Thus, the STFT computed with an N-point Hamming window needs only be evaluated at intervals of $R = N/4$. Because the Fourier transform of a signal of length N can be reconstructed from N frequency samples, the number of time and frequency samples that are necessary to represent the STFT is only 4 times greater than the number of samples in the signal $x[n]$. In fact, for real signals the number of STFT samples needs only be twice the number of signal samples because of the conjugate symmetry of Fourier transforms. In many applications, this slight redundancy of the STFT is more than offset by increased flexibility in processing. Furthermore, it is often possible to sample the STFT at a much lower rate than $4/N$, while preserving essential features of the signal (such as speech intelligibility). This observation is the basis for *vocoders,* systems for the efficient transmission of speech.

Because the FFT gives equally-spaced frequency samples, this implementation is most useful when all the filters in STFT the filter bank have the same bandwidth. In some cases, unequal bandwidths are desirable in short-time Fourier analysis. For example when modeling cochlear processing, filter bandwidths are approximately a constant fraction of their center frequencies. Because filter bandwidths are inversely proportional to the duration of the impulse response, this implies that window lengths should be inversely proportional to filter center frequencies. Such proportional scaling can be efficiently achieved using the *Wavelet transform.* This transform gives good frequency resolution at low frequencies, and good time resolution at high frequencies.

7.3 Applications of the short-time Fourier transform

7.3.1 Spectrographic displays

A *spectrogram* is a display of the magnitude of the short-time Fourier transform of a signal as a function of both time and frequency. For sound signals, restriction to the magnitude of the STFT is usually justified because the ear is not very sensitive to the phase of the short-time spectrum (This is known as *Ohm's acoustic law).* Spectrograms can be displayed either by encoding energy on a gray scale, or as perspective (”waterfall”) representations. Gray-scale displays are particularly convenient, and were produced by analog spectrographic instruments in the 1940's, long before digital spectrograms became available. Analog spectrograms were generated by passing the signal through a bank of analog bandpass filters, then computing the short-time energy at the output of each filter by rectifying and lowpass filtering the bandpass filter outputs (bandpass filter interpretation of the STFT). Modern, digital spectrograms are obtained by computing fast Fourier transforms of successive signal segments as described above (Fourier transform interpretation of the STFT).

Two bandwidths are widely used in spectrographic analyses of speech: *Broadband* spectrograms, which have a frequency resolution of 300 Hz, and *narrowband* spectrograms which have a resolution of 50 Hz. Broadband and narrowband spectrograms of an utterance produced by a male speaker are shown in Fig. 9. The frequency resolution of the narrowband spectrogram is sufficient to resolve individual harmonics of the fundamental frequency of voice (˜100 Hz). These harmonics appear as horizontal bands during voiced portions of speech. On the other hand, the broadband spectrogram has sufficient time resolution to resolve individual opening and closing of the vocal cords, which appear as vertical striations during voiced segments. Thus, the periodic vibration of the vocal cords appears as vertical striations in broadband spectrograms, and as horizontal bands in narrowband spectrograms. The broadband spectrogram also reveals the short noise bursts of stop consonants and rapid changes in formant frequencies.

7.3.2 Time-dependent filtering

The STFT is useful not only for analyzing nonstationary signals, but also for carrying out time-varying modifications of signals. For example, to implement a time-dependent filter, the STFT of the input signal $x[n]$ is multiplied by a time-dependent frequency response $H_n(f)$, then the inverse STFT of the product $X_n(f)\ H_n(f)$ is computed. Other applications of the STFT are to translate or scale the signal spectrum, and to speed up or slow down signals without changing their frequency content. Just as speeding up a tape recorder by a factor K scales up the frequency axis by a factor of K, playing back a sampled signal at a different sampling frequency also scales its spectrum. Appropriate use of the STFT can compensate for this spectral scaling. In these applications, it is necessary to synthesize a signal from the STFT, i.e. to have an inverse STFT algorithm. Such an algorithm is described in the Appendix.

7.3.3 Analysis-synthesis systems

An *analysis-synthesis system (or "vocoder")* is a device for efficient transmission or storage of speech or audio signals. A measure of efficiency is the *bit rate,* i.e. the number of bits per second that are required to represent the signal. For example, a high-quality audio signal is typically sampled at 48 kHz and quantized at 16 bits, requiring a rate of 1.5 M bits/s for a two-channel (stereo) signal. A telephone-quality speech signal sampled at 8 kHz with 10-bit quantization requires a rate of 80,000 bits/s. Vocoders allow the bit rate to be considerably decreased (to as little as 1200 - 9600 bits/s) with little degradation in speech intelligibility. A vocoder consists of three stages: an *analysis stage* which extracts information-bearing parameters from the signal and encodes them for transmission or storage, an optional *transmission channel* or storage medium, and a *synthesis stage* which decodes the transmitted parameters and generates a signal from these parameters.

The STFT is the basis for the *channel vocoder,* which was developed by Homer Dudley at Bell Laboratories in the 1930's, and is still a widely used analysis-synthesis system, particularly in applications with high background noise. The channel vocoder is based on the *source-filter model* of speech production. This model states that speech can be reconstructed from two sets of signals, one representing a source and one representing the filtering properties of the vocal organs. The source can be either a *voicing source* representing periodic vibration of the vocal cords, or a *noise source* representing turbulence generated in the vocal tract. In a channel vocoder (Fig. 10), the analysis stage extracts filter information from the magnitude of the STFT evaluated at a small number of frequency channels (typically 10-20). Again, phase information can be discarded because the ear is not particularly sensitive to phase. Source information is obtained by a pitch/voicing detector, which determines if each frame is voiced, and, if so, what is the fundamental frequency of the vocal cords (which is heard as pitch). After analysis, voicing and pitch information are be transmitted together with the STFT magnitude for each frequency channel. At the synthesis stage, a speech signal is reconstructed by exciting the filters by a source signal whose amplitude for each channel is modulated by the transmitted STFT magnitude. Either of two source signals is used: a periodic train of pulses for voiced segments, and broadband noise during unvoiced segments. The frequency of the pulse train is determined by the detected pitch. In practice, 10-20 frequency channels, each sampled at intervals of 10 ms suffice to produce intelligible speech. If the energy in each channel is encoded with 6 bits, this gives a rate of 9000 bits/sec, roughly a factor of 10 decrease over pulse-code modulation.

In recent years, another type of analysis-synthesis system based on the STFT has been used in high-quality digital processing of sound signals, particularly music. Unlike the channel vocoder, these new systems require transmission of both the magnitude and phase of the STFT. These systems make use of knowledge of masking properties of the human ear to reduce the bit rate necessary to encode the STFT. For example, if at a particular time, the sound signal contains both a loud frequency component and a weak component, the weak component will be masked (i.e. not heard) if it is sufficiently close in frequency to the loud component. Under such conditions, there would be no loss in quality if the weak component were removed (i.e. not transmitted) in the analysis-synthesis system. Using these ideas, compact-disk quality encoding of stereo sound has been demonstrated with bit rates of 384 kB/sec, about 1/4 of the rate used in CD players.

7.4 Summary

Time-dependent signal processing consists in making certain measurements for successive short-time segments of frasmes of a nonstationary signal. It is a useful technique for signals whose statistical characteristics change slowly relative to the durations of the analysis frames. Virtually any form of processing that is applicable to stationary signals can also be carried out for short time segments of nonstationary signals: Examples include the short-time energy, the short-time autocorrelation function, and the short-time Fourier transform.

Short-time Fourier analysis is used for representing the energy distribution of nonstationary signals as a function of both time and frequency. The short-time Fourier transform can be implemented either by computing the Fourier transforms of successive, windowed signal segments, or by passing the signal through a bank of bandpass filters. The time resolution of this analysis is determined by the duration of the window, while the frequency resolution is determined by the passband of the bandpass filters, which is equal to the bandwidth of the window. Because the uncertainty principle places a lower bound on the duration-bandwidth product of any signal, it is not possible to simultaneously achieve fine time resolution and fine frequency resolution in short-time Fourier analysis.

7.5 Further reading

Rabiner and Schafer, Chapter 4, Sections 1, 2, 6; Chapter 6, Sections 1, 4
Quatieri, Chapter 3, Section 3; Chapter 7, Sections 1-4

7.6 Appendix: Signal reconstruction from the short-time Fourier transform

Equation (7.9) shows that exact signal reconstruction from the STFT is theoretically possible. However, this formula is not practically useful because it requires knowing the STFT for all frequencies. We will show that exact reconstruction is possible when a finite number of *frequency samples* of the STFT are available, providing that the analysis window verifies a simple condition. To be specific, assume that the STFT $X_n(f)$ is available for N frequency samples $f_k = k/N$, $0 \leq k \leq N-1$, and for all n. We propose to reconstruct the signal by approximating the integral (7.9) by a sum over the frequency samples:

$$\hat{x}[n] \triangleq \frac{1}{N\ w[0]} \sum_{k=0}^{N-1} X_n(f_k)\ e^{j2\pi f_k n} = \frac{1}{N\ w[0]} \sum_{k=0}^{N-1} \tilde{X}_n(f_k) \qquad (7.A.1)$$

This *filter-bank summation* method of synthesis is illustrated in Fig. A.1. The figure shows that the reconstructed signal $\hat{x}[n]$ can be considered as the response of a parallel combination of N bandpass filters $w_{f_k}[n]$ to the signal $x[n]$. Therefore, the reconstructed signal will be exactly equal to the original signal if and only if the frequency response of the parallel combination of filters is 1 for all frequencies:

$$\frac{1}{N\ w[0]} \sum_{k=0}^{N-1} W_{f_k}(f) = 1 \qquad (7.A.2)$$

We have seen that the DTFT of $w_{f_k}[n] = w[n]\ e^{j2\pi f_k n}$ is $W(f - f_k)$. Thus, replacing f_k by its value k/N, the condition for exact reconstruction becomes:

$$\frac{1}{N\ w[0]} \sum_{k=0}^{N-1} W(f - k/N) = 1 \qquad (7.A.3)$$

For example, Figure A.2a shows that this condition is verified if $W(f)$ is an ideal lowpass filter with cutoff frequency $1/2N$.

While the ideal lowpass window is of theoretical interest, it cannot be realized by a finite digital filter. More practical windows can be found by writing the condition for exact reconstruction (7.A.3) in the time domain. For this purpose, we note that $W(f - k/N)$ is equal to the cyclic convolution $W(f) \circledast \tilde{\delta}(f - k/N)$. Reporting this expression into (7.A.3), and making use of the distributivity of convolution with respect to addition yields:

$$\frac{1}{N\ w[0]} \sum_{k=0}^{N-1} W(f) \circledast \tilde{\delta}(f - k/N) = \frac{1}{N\ w[0]}\ W(f) \circledast \sum_{k=0}^{N-1} \tilde{\delta}(f - k/N) = 1$$

Using the product theorem and the Fourier transform pair for periodic impulse trains, we obtain:

$$w[n] \sum_{r=-\infty}^{\infty} \delta(n - rN) \longleftrightarrow 1/N \; W(f) \circledast \sum_{l=-\infty}^{\infty} \delta(f - k/N) \;=\; \frac{1}{N \; w[0]} \; W(f) \circledast \sum_{k=0}^{N-1} \tilde{\delta}(f - k/N)$$

Noting that the inverse DTFT of 1 is $\delta[n]$, we obtain the condition for exact reconstruction in the time domain:

$$\sum_{r=-\infty}^{\infty} w[rN] \; \delta[n - rN] = \; w[0] \; \delta[n] \tag{7.A.4}$$

This condition simply requires that $w[n]$ be zero for all times n that are non-zero multiples of the number of frequency samples N. Figure A.2 shows examples of windows that verify this condition. Note that, to verify (7.A.4), it suffices that the duration of $w[n]$ be less than N, but that this condition is by no means necessary.

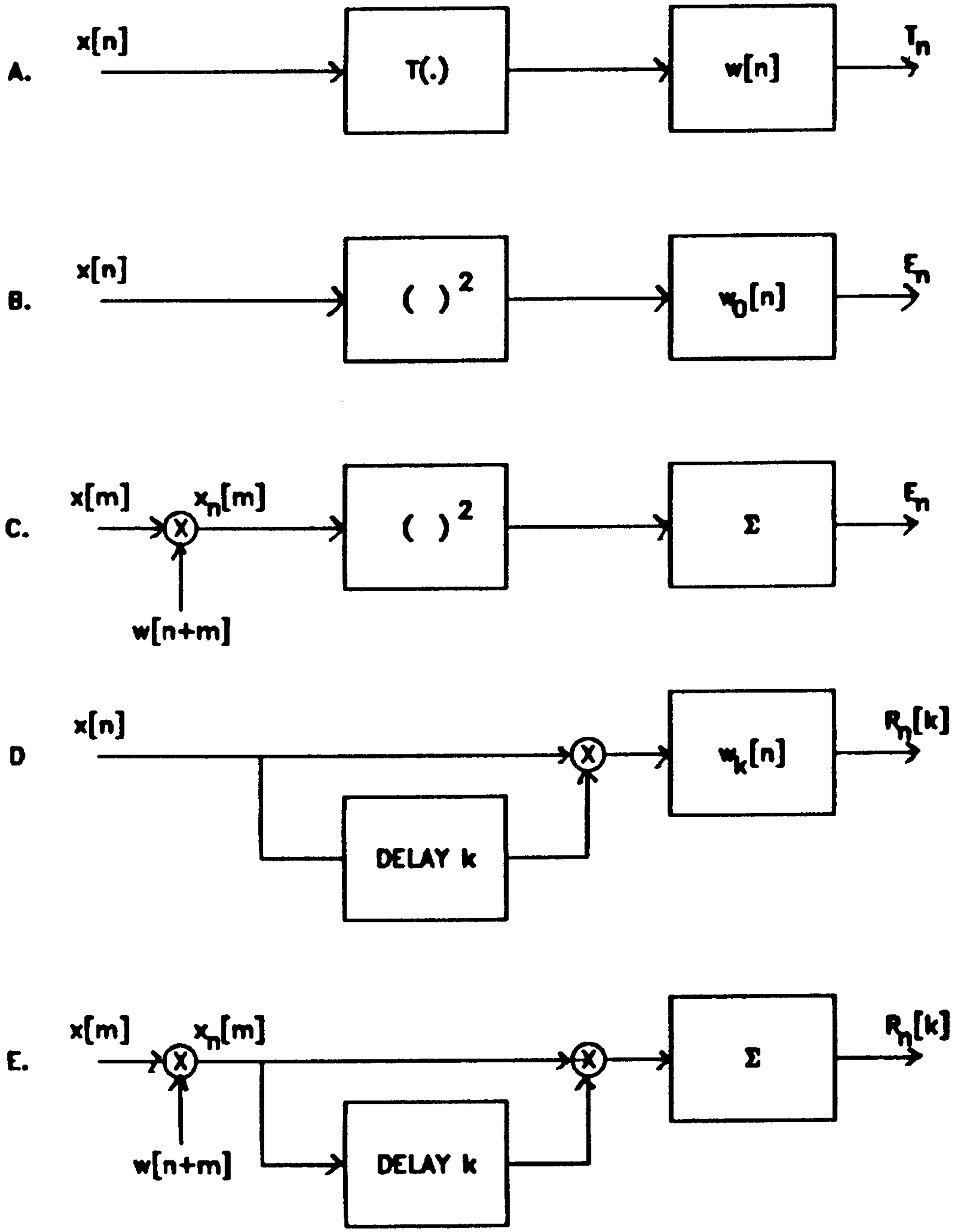

Figure 7.1: Block-diagram representation of various types of short-time processing. A. General time-dependent measurement. B. Low-pass filter implementation of the short-time energy. C. Window implemetation of the short-time energy. D. Low-pass filter implementation of the short-time autocorrelation function. E. Window implementation of the short-time autocorrelation function.

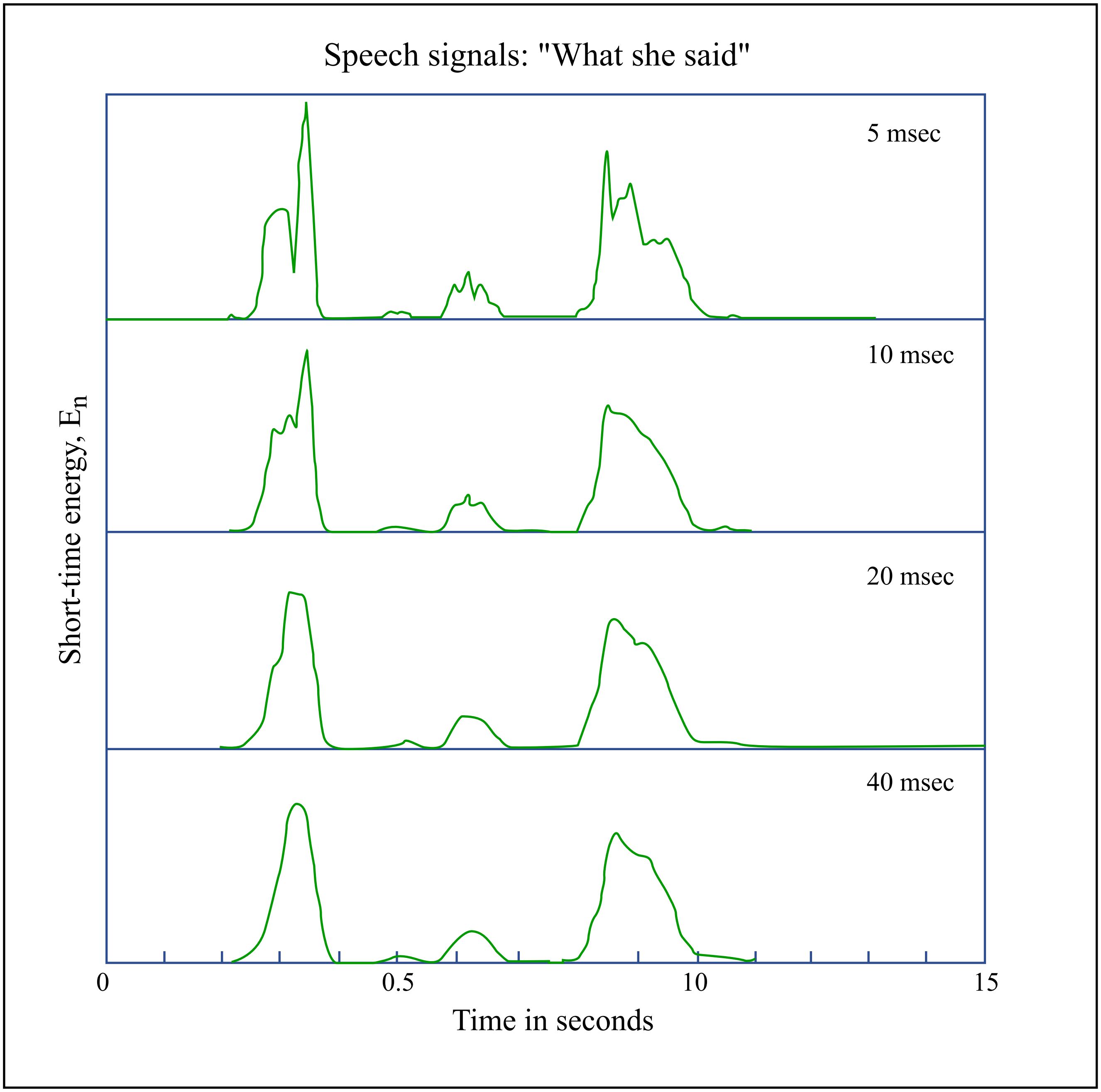

Figure by MIT OpenCourseWare. After Rabiner and Schafer.

Figure 7.2: Short-time energy functions of a speech signal for Ham- ming windows of 5, 10, 20 and 40 ms.

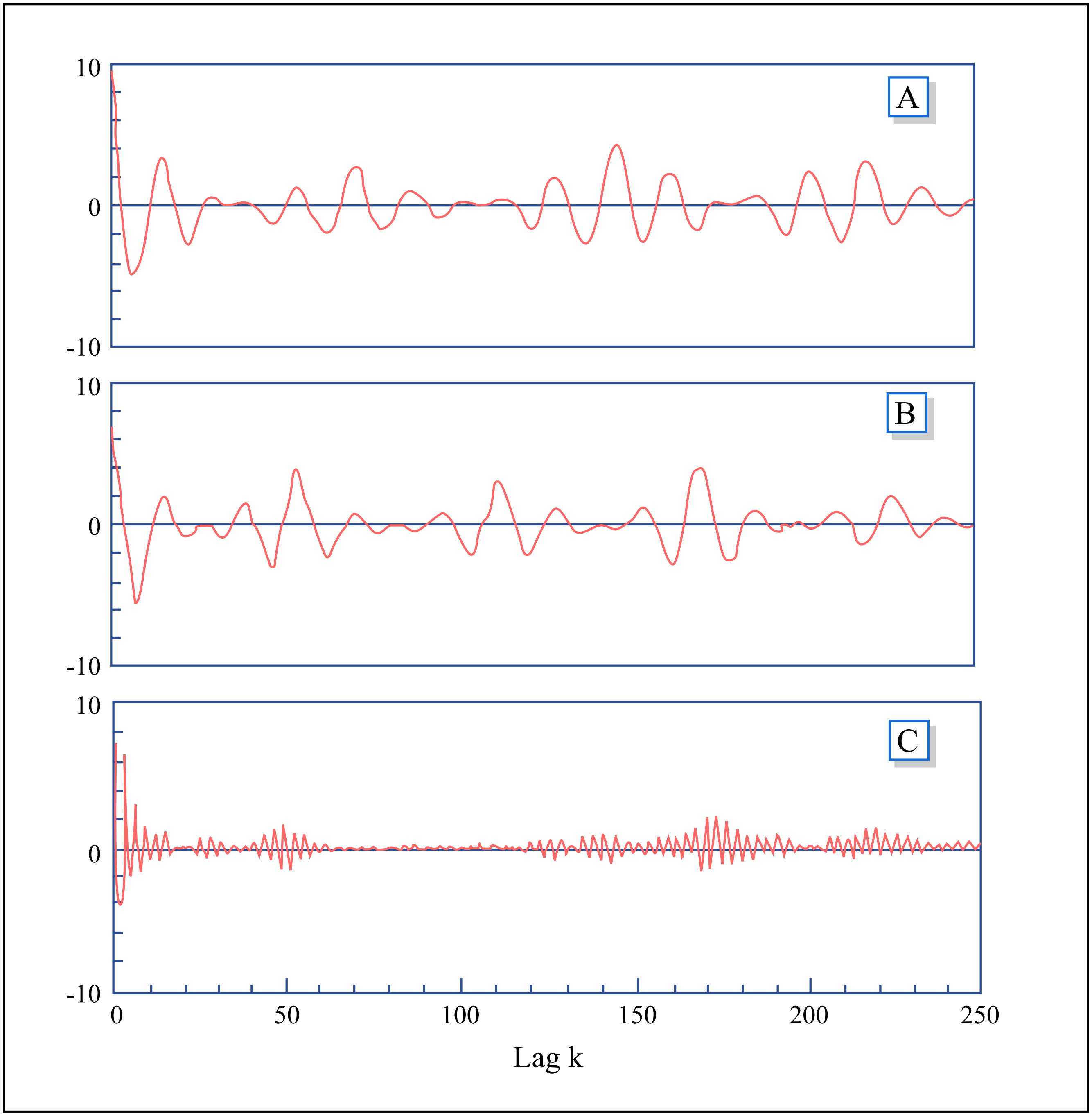

Figure by MIT OpenCourseWare. After Rabiner and Schafer.

Figure 7.3: Short-time autocorrelation functions for (a) and (b) voiced speech, and (c) unvoiced speech using 40-ms Hamming windows.

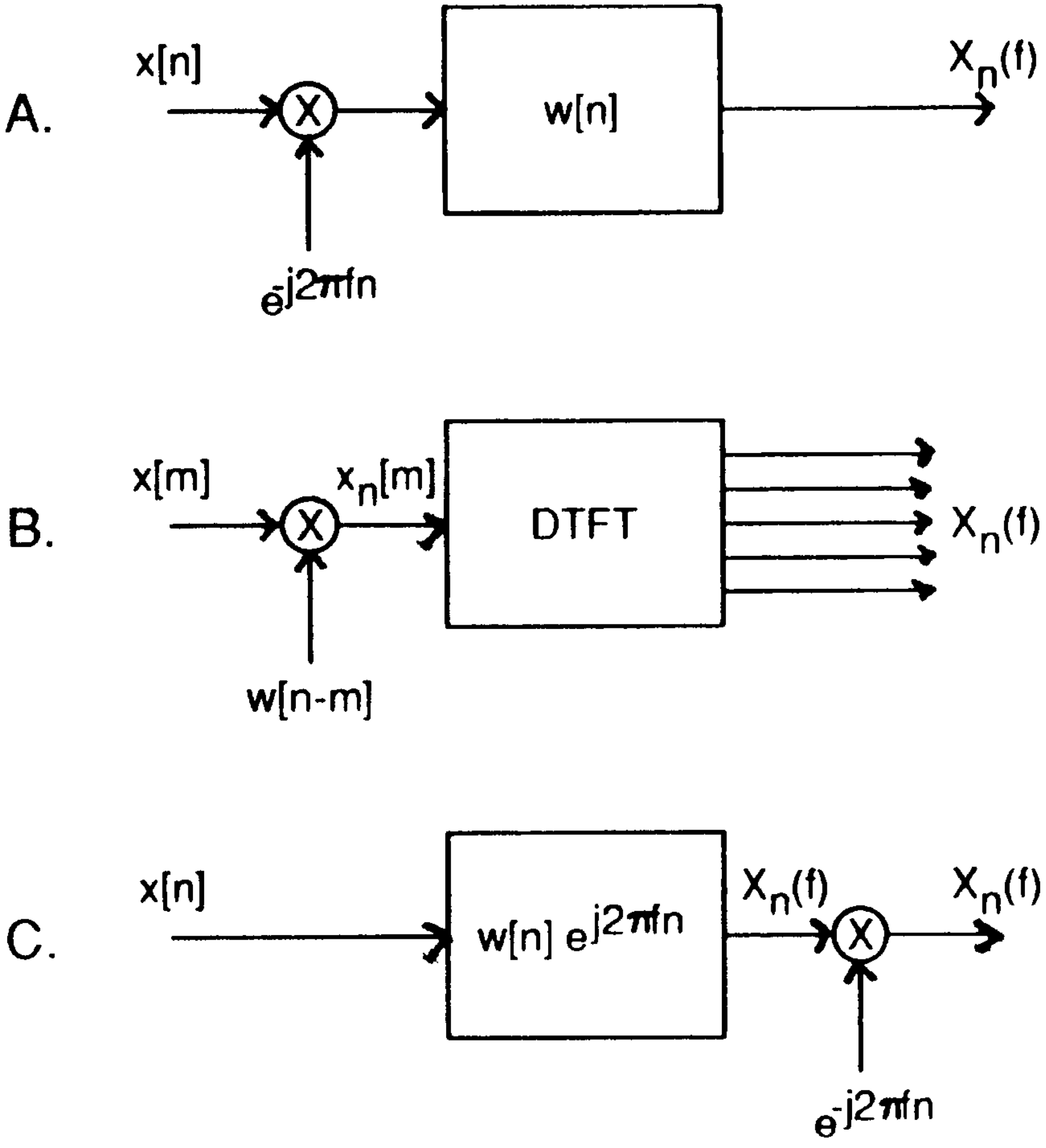

Figure 7.4: A. Time-dependent processing interpretation of the STFT. B. Fourier transform interpretation of the STFT. C. Bandpass filtering interpretation of the STFT.

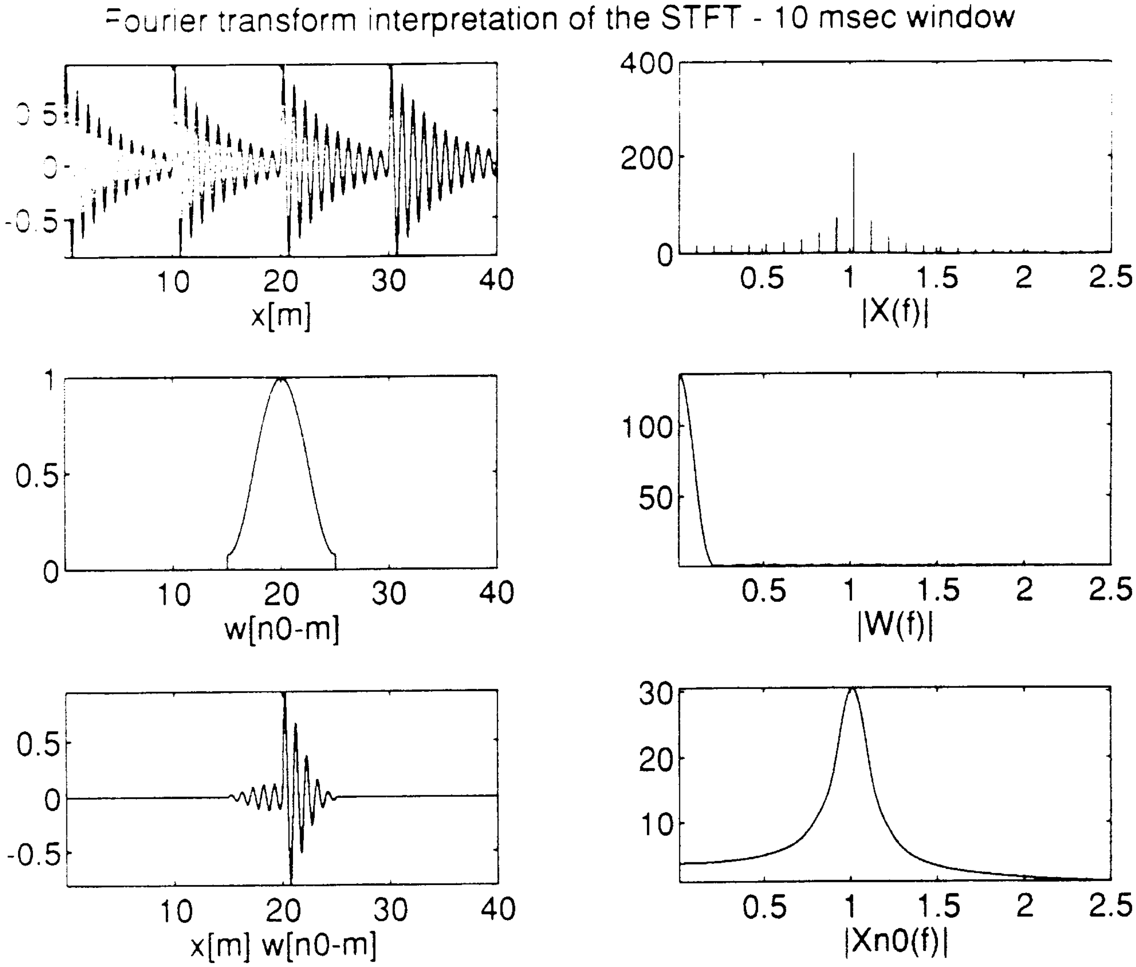

Figure 7.5: Fourier transform interpretation of the STFT - 10 msec window.

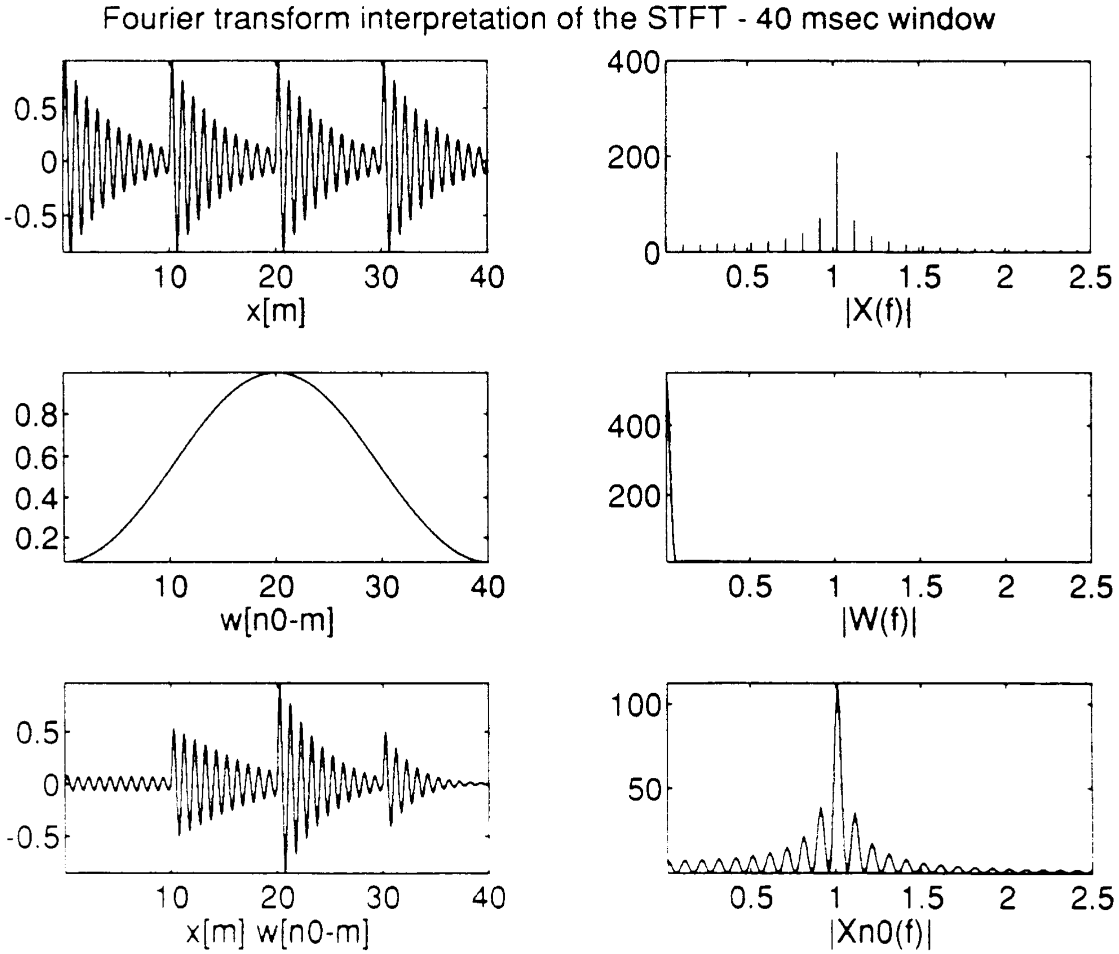

Figure 7.6: Fourier transform interpretation of the STFT - 40 msec window.

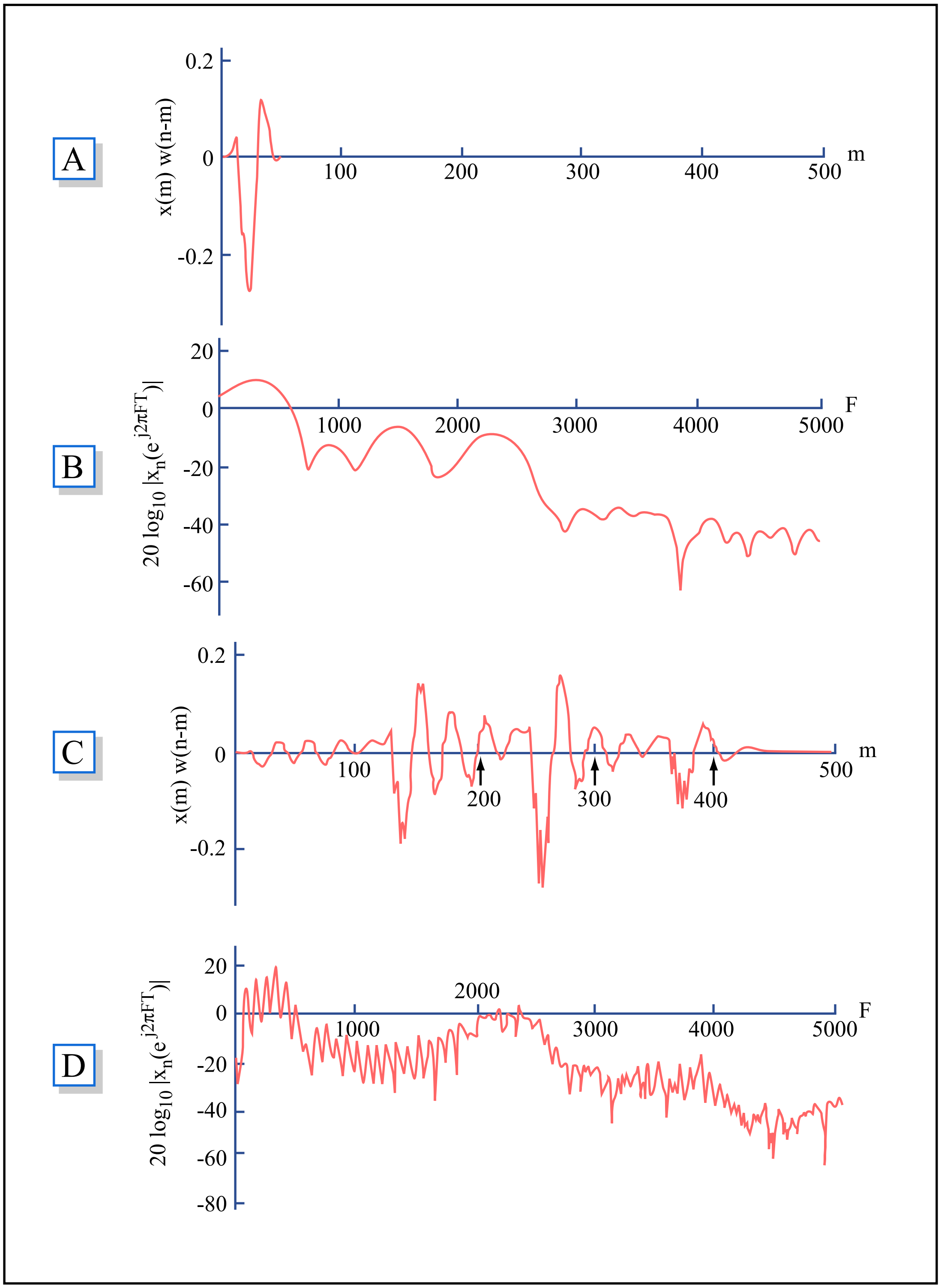

Figure by MIT OpenCourseWare. After Rabiner and Schafer.

Figure 7.7: Waveforms and short-time spectra of vowels obtained with 5-ms (top) and 50-ms (bottom) Hamming windows.

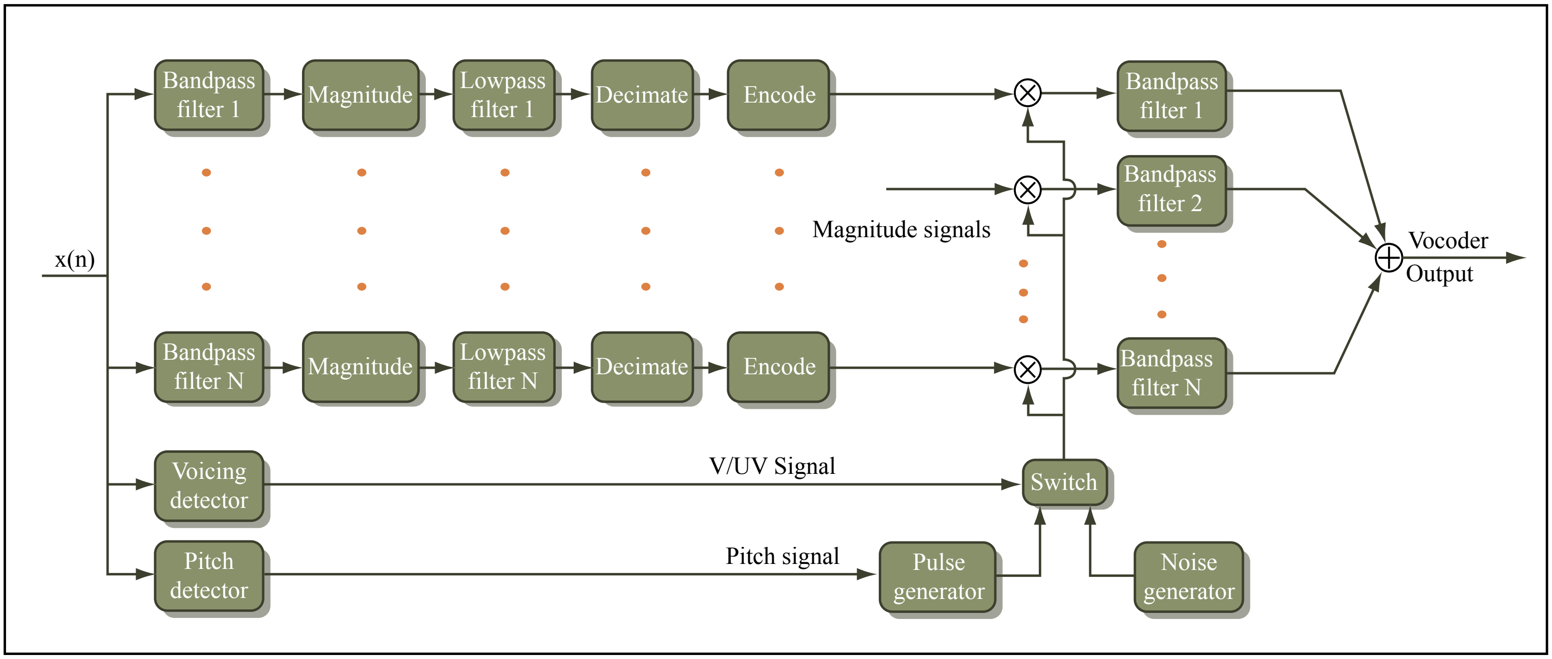

Figure by MIT OpenCourseWare. After Rabiner and Schafer.

Figure 7.8: Block diagram of a channel vocoder. Top: analysis stage.

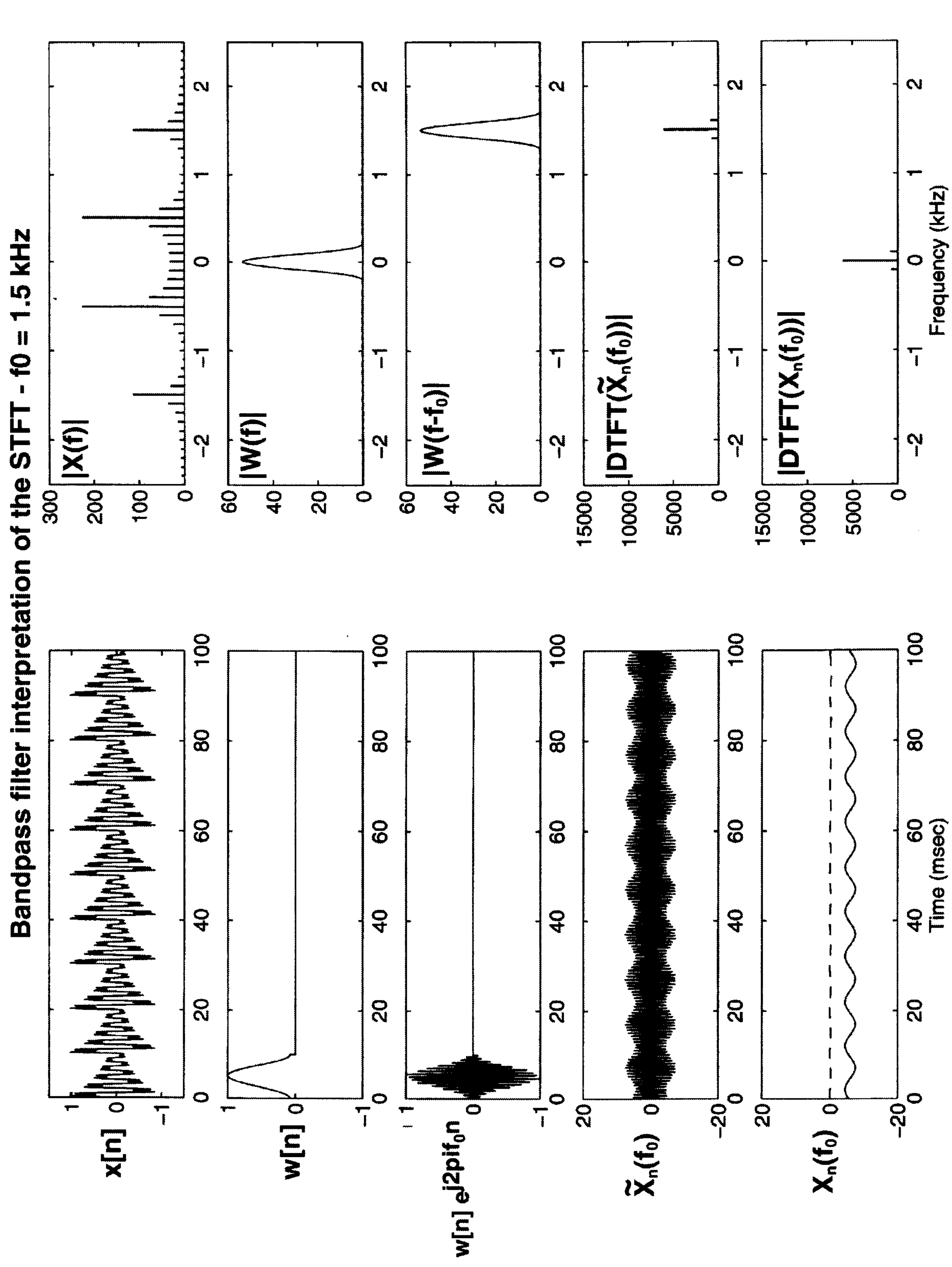

Figure 7.9: Bandpass filter interpretation of the STFT - $f_0 = 1.5kHz$.

Image removed due to copyright restrictions.

Please see: Figure 10.18 in Oppenheim and Schafer. *Discrete-Time Signal Processing*. 1st ed. Upper Saddle River: Prentice-Hall, 1975.

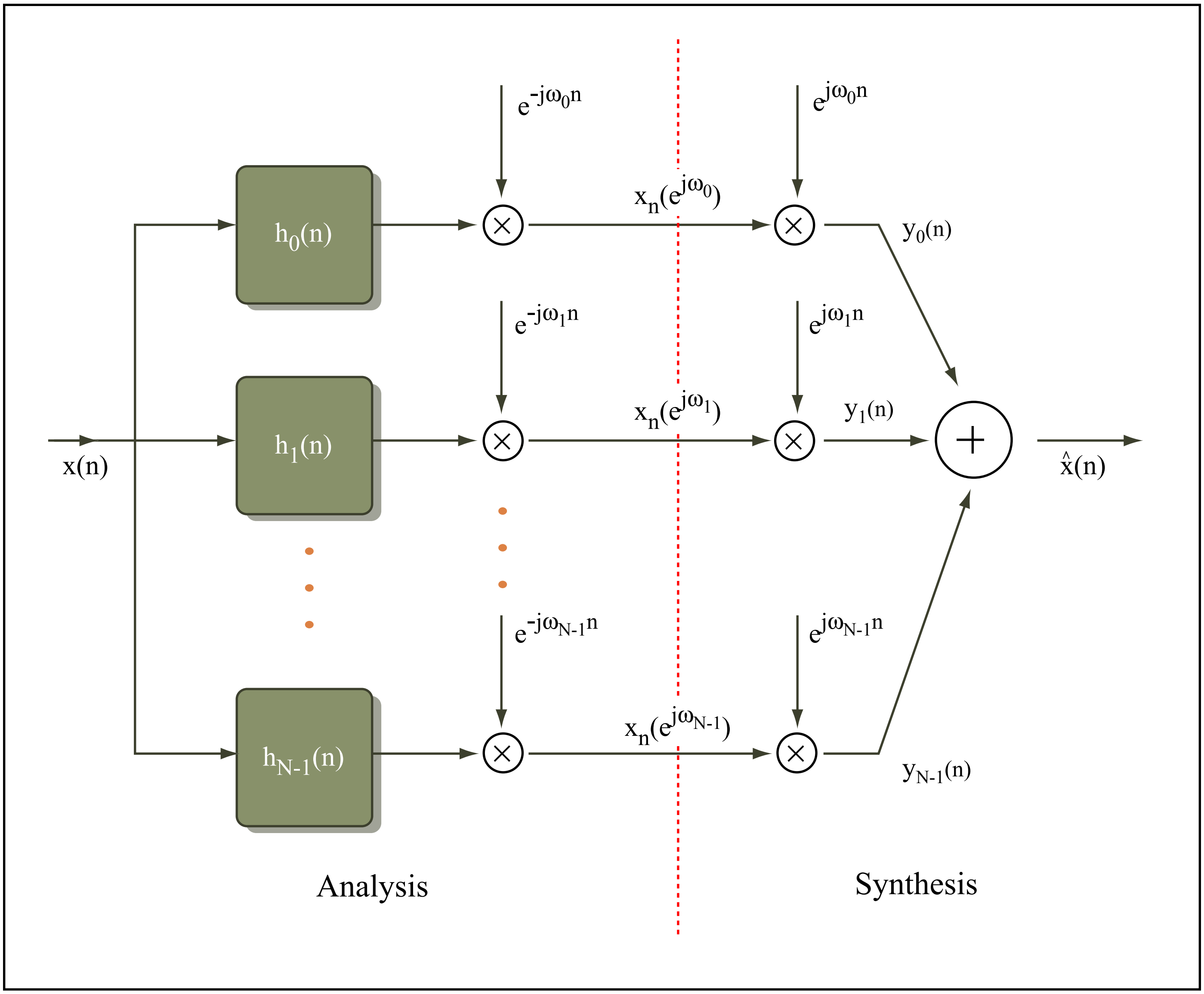

Figure by MIT OpenCourseWare. After Rabiner and Schafer.

Figure 7.11: Signal analysis and synthesis be the filter bank sum- mation method.

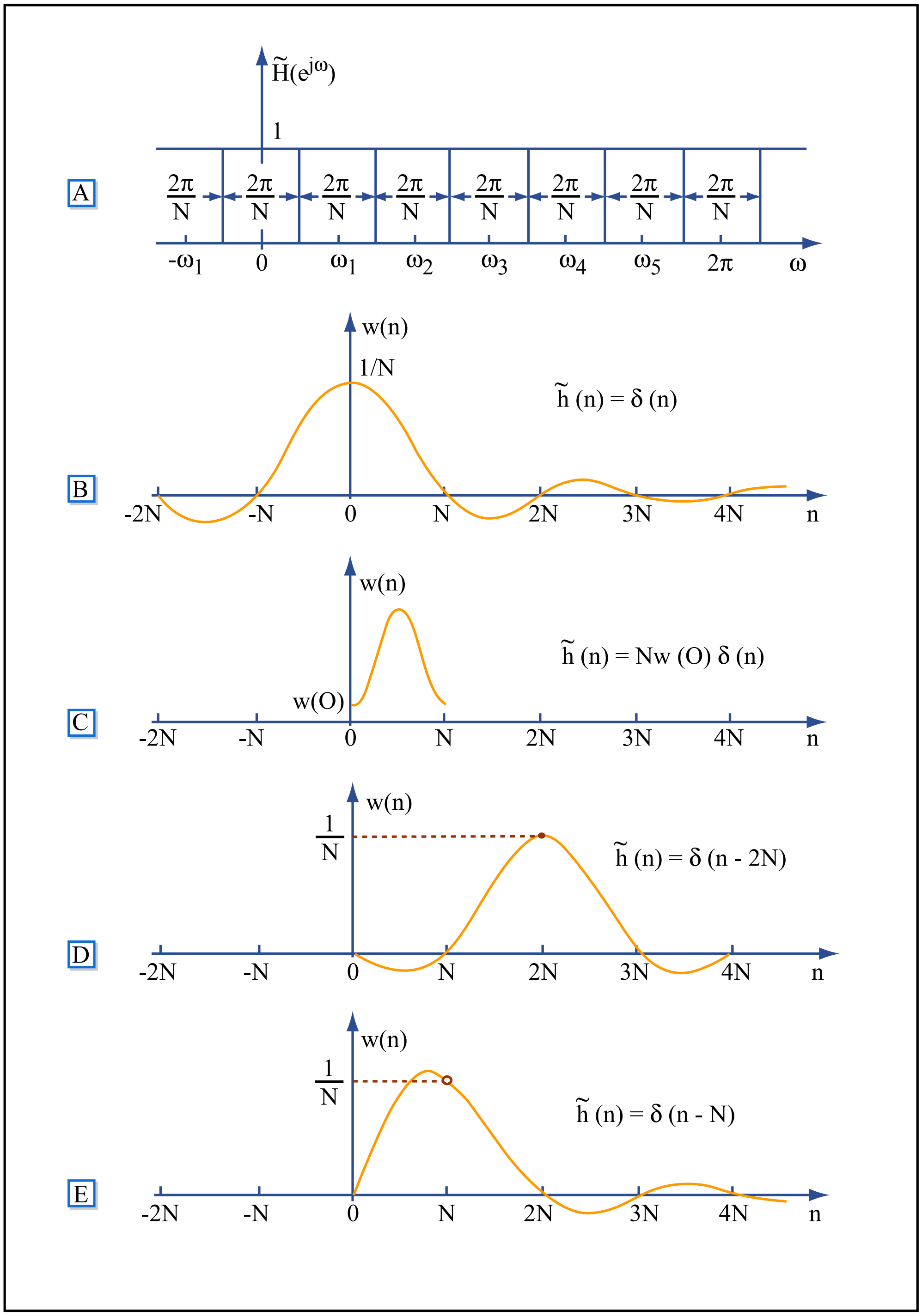

Figure by MIT OpenCourseWare. After Rabiner and Schafer.

Figure 7.12: A. Composite frequency response of the filter bank summation method of signal synthesis for $N = 6$ ideal filters. B. Typical windows that verify the exact reconstruction condition in the filter bank summation method of signal synthesis.