For windowing of voiced speech, a rectangular window with a duration of one pitch period (and centered on the period) produces an output spectrum close to that of the vocal tract impulse response, to the extent that each pitch period corresponds to such an impulse response. (This works best for low-F0 voices, where the pitch period is long enough to permit the signal to decay to low amplitude before the next vocal cord closure.) Unfortunately, it is often difficult to reliably locate pitch periods for such *pitch-synchronous* analysis, and system complexity increases if window size must change dynamically with F0. Furthermore, since most pitch periods are indeed shorter than the vocal tract impulse response, a one-period window truncates, resulting in spectral degradation.

For simplicity, most speech analyses use a fixed window size of longer duration, e.g., 25 ms. Problems of edge effects are reduced with longer windows; if the window is shifted in time without regard for pitch periods in the common *pitch-asynchronous* analysis, the more periods under the window the less the effects of including/excluding the large-amplitude beginning of any individual period. Windows well exceeding 25 ms smooth rapid spectral changes (relevant in most applications) too much. For F0 estimation, however, windows must typically contain at least two pitch periods; so pitch analysis uses a long window—often 30–50 ms.

Recent attempts to address the drawbacks of a fixed window size include more advanced frequency transforms (e.g., wavelets—see below), as well as simpler modifications to the basic DFT approach (e.g., the 'modulation spectrogram' [1], which emphasizes slowly varying speech changes around 4 Hz, corresponding to approximate syllable rates, at the expense of showing less rapid detail).

6.3 TIME-DOMAIN PARAMETERS

Analyzing speech in the time domain has the advantage of simplicity in calculation and physical interpretation. Several speech features relevant for coding and recognition occur in temporal analysis, e.g., energy (or amplitude), voicing, and F0. Energy can be used to segment speech in automatic recognition systems, and must be replicated in synthesizing speech; accurate voicing and F0 estimation are crucial for many speech coders. Other time features, e.g., zero-crossing rate and autocorrelation, provide inexpensive spectral detail without formal spectral techniques.

6.3.1 Signal Analysis in the Time Domain

Time-domain analysis transforms a speech signal into a set of parameter signals, which usually vary much more slowly in time than the original signal. This allows more efficient storage or manipulation of relevant speech parameters than with the original signal; e.g., speech is usually sampled at 6000–10,000 samples/s (to preserve bandwidth up to 3–5 kHz), and thus a typical 100 ms vowel needs up to 1000 samples for accurate representation. The information in a vowel relevant to most speech applications can be represented much more efficiently: energy, F0, and formants usually change slowly during a vowel. A parameter signal at 40–100 samples/s suffices in most cases (although 200 samples/s could be needed to accurately track rapid changes such as stop bursts). Thus, converting a speech waveform into a set of parameters can decrease sampling rates by two orders of magnitude. Capturing the relevant aspects of speech, however, requires several parameters sampled at the lower rate.

While time-domain parameters alone are rarely adequate for most applications, a combined total of 5–15 time- and frequency-domain parameters often suffice.

Most short-time processing techniques (in both time and frequency) produce parameter signals of the form

$$Q(n) = \sum_{m=-\infty}^{\infty} T[s(m)]w(n-m). \tag{6.4}$$

The speech signal $s(n)$ undergoes a (possibly nonlinear) transformation T, is weighted by the window $w(n)$, and is summed to yield $Q(n)$ at the original sampling rate, which represents some speech property (corresponding to T) averaged over the window duration. $Q(n)$ corresponds to a convolution of $T[s(n)]$ with $w(n)$. To the extent that $w(n)$ represents a lowpass filter, $Q(n)$ is a smoothed version of $T[s(n)]$.

Since $Q(n)$ is the output of a lowpass filter (the window) in most cases, its bandwidth matches that of $w(n)$. For efficient manipulation and storage, $Q(n)$ may be decimated by a factor equal to the ratio of the original sampled speech bandwidth and that of the window; e.g., a 20 ms window with an approximate bandwidth of 50 Hz allows sampling of $Q(n)$ at 100 samples/s (100:1 decimation if the original rate was 10,000 samples/s). As in most decimation operations, it is unnecessary to calculate the entire $Q(n)$ signal; for the example above, $Q(n)$ need be calculated only every 10 ms, shifting the analysis window 10 ms each time. For any signal $Q(n)$, this eliminates much (mostly redundant) information in the original signal. The remaining information is in an efficient form for many speech applications.

In addition to the common rectangular and Hamming windows, the Bartlett, Blackman, Hann, Parzen, or Kaiser windows [2, 3] are used to smooth aspects of speech signals, offering good approximations to lowpass filters while limiting window duration (see Figure 6.2). Most windows have finite-duration impulse responses (FIR) to strictly limit the analysis time range, to allow a discrete Fourier transform (DFT) of the windowed speech and to preserve phase. An infinite-duration impulse response (IIR) filter is also practical if its z transform is a rational function; e.g., a simple IIR filter with one pole at $z = a$ yields a recursion:

$$Q(n) = aQ(n-1) + T[s(n)]. \tag{6.5}$$

IIR windows typically need less computation than FIR windows, but $Q(n)$ must be calculated at the original (high) sampling rate before decimating. (In real-time applications, a speech measure may be required at every sample instant anyway). FIR filters, having no recursive feedback, permit calculation of $Q(n)$ only for the desired samples at the low decimated rate. Most FIR windows of N samples are symmetric in time; thus $w(n)$ has linear phase with a fixed delay of $(N-1)/2$ samples. IIR filters do not permit simple delay compensation.

6.3.2 Short-Time Average Energy and Magnitude

$Q(n)$ corresponds to short-time energy or amplitude if T in Equation (6.4) is a squaring or absolute magnitude operation, respectively (Figure 6.5). Energy emphasizes high amplitudes (since the signal is squared in calculating $Q(n)$), while the amplitude or magnitude measure avoids such emphasis and is simpler to calculate (e.g., with fixed-point arithmetic, where the dynamic range must be limited to avoid overflow). Such measures can help segment speech into smaller phonetic units, e.g., approximately corresponding to syllables or phonemes. The large variation in amplitude between voiced and unvoiced speech, as well as smaller variations between phonemes with different manners of articulation, permit segmentations based on energy $Q(n)$ in automatic recognition systems. For isolated word recognition,

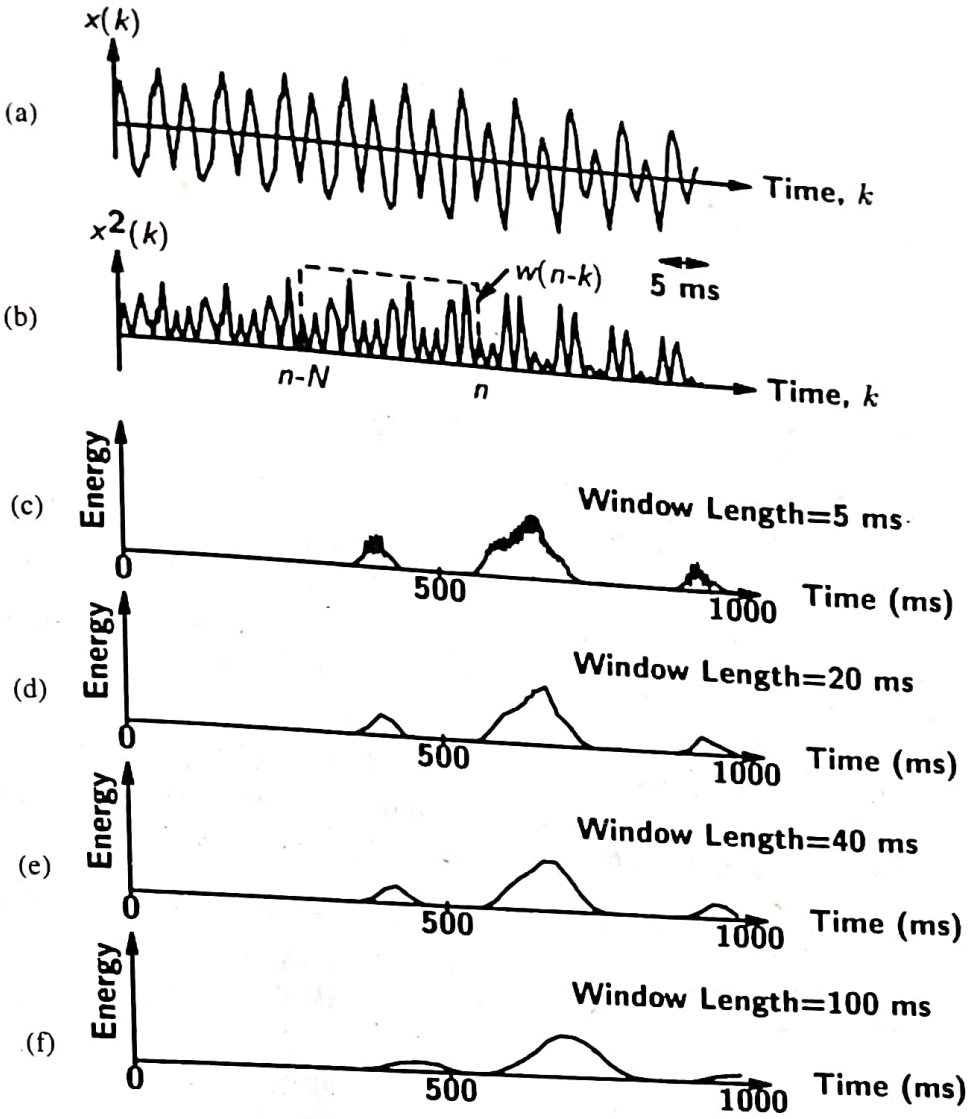

Figure 6.5 Illustration of the computation of short-time energy: (a) 50 ms of a vowel, (b) the squared version of (a), with a superimposed window of length N samples delayed n samples, (c–f) energy function for a 1 s utterance, using rectangular windows of different lengths.

such $Q(n)$ can aid in accurate determination of the endpoints of a word surrounded by pauses. In speech transmission systems that multiplex several conversations, this $Q(n)$ can help detect the boundaries of speech, so that pauses need not be sent.

6.3.3 Short-Time Average Zero-crossing Rate (ZCR)

Normally, spectral measures of speech require a Fourier or other frequency transformation or a complex spectral estimation (e.g., linear prediction). For some applications, a simple measure called the zero-crossing rate (ZCR) provides adequate spectral information at low cost. In a signal $s(n)$ such as speech, a *zero-crossing* occurs when $s(n) = 0$, i.e., the waveform crosses the time axis or changes algebraic sign. For narrowband signals (e.g., sinusoids), ZCR (in zero-crossings/s) is an accurate spectral measure; a sinusoidal has two zero-crossings/period, and thus its F0 = ZCR/2.

For discrete-time signals with ZCR in zero-crossings/sample,

$$\text{F0} = (\text{ZCR} * F_s)/2, \tag{6.6}$$

for F_s sample/s.

The ZCR can be defined as $Q(n)$ in Equation (6.4), with

$$T[s(n)] = 0.5|\text{sgn}(s(n)) - \text{sgn}(s(n-1))| \tag{6.7}$$

where the algebraic sign of $s(n)$ is

$$\text{sgn}(s(n)) = \begin{cases} 1 & \text{for } s(n) \geq 0 \\ -1 & \text{otherwise,} \end{cases} \tag{6.8}$$

and $w(n)$ is a rectangular window scaled by $1/N$ (where N is the duration of the window) to yield zero-crossings/sample, or by F_s/N to yield zero-crossings/s. This $Q(n)$ can be heavily decimated since the ZCR changes relatively slowly with the vocal tract movements.

The ZCR can help in voicing decisions. Most energy in voiced speech is at low frequency, since the spectrum of voiced glottal excitation decays at about −12 dB/oct. In unvoiced sounds, broadband noise excitation excites mostly higher frequencies, due to effectively shorter vocal tracts. While speech is not a narrowband signal (and thus the sinusoid example above does not hold), the ZCR correlates well with the average frequency of major energy concentration. Thus high and low ZCR correspond to unvoiced and voiced speech, respectively. A suggested boundary is 2500 crossings/s, since voiced and unvoiced speech average about 1400 and 4900 crossings/s, respectively, with a larger standard deviation for the latter (Figure 6.6).

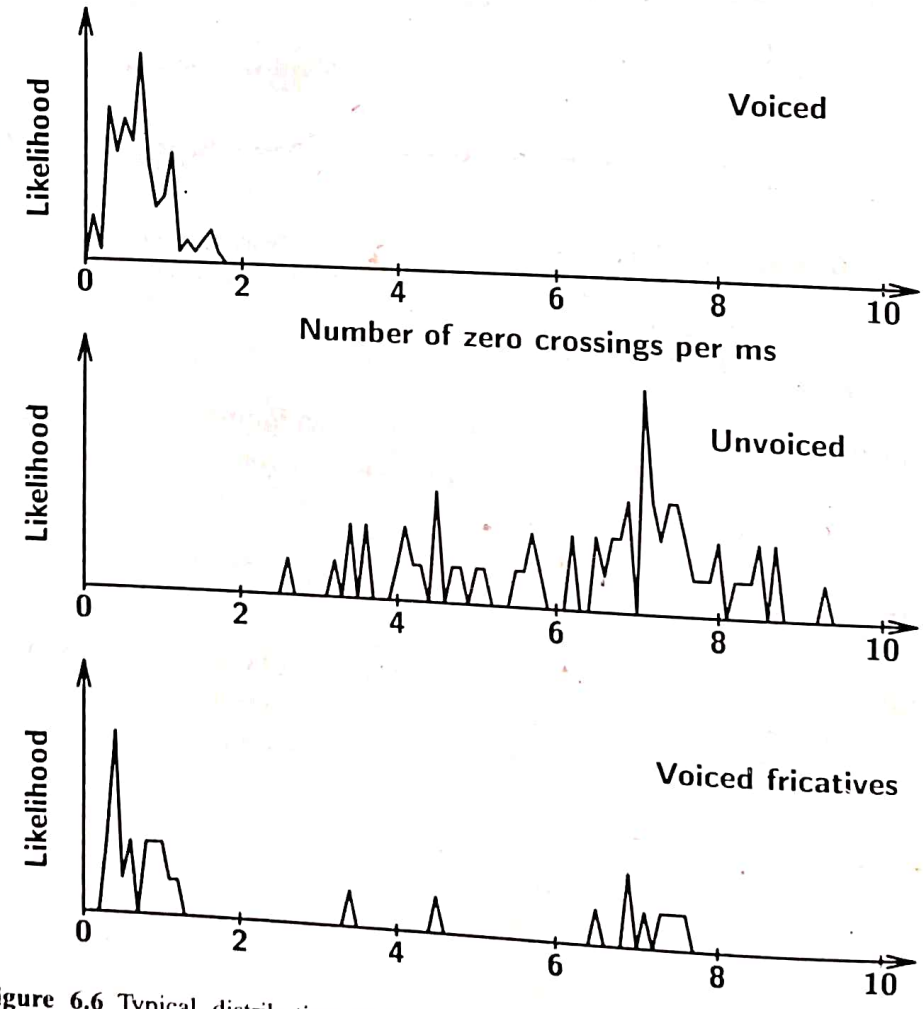

Figure 6.6 Typical distribution of zero-crossings for voiced sonorants, for unvoiced frication, and for voiced frication.

For vowels and sonorants, the ZCR corresponds mostly to F1, which has more energy than other formants. Interpreting ZCR is harder for voiced fricatives, which have both periodic energy in the voice bar at very low frequency and unvoiced energy at high frequency. This, of course, is a problem for all voiced/unvoiced determination methods; a binary decision using a simple threshold test on the ZCR is inadequate. Depending on the balance of periodic and aperiodic energy in voiced fricatives, some are above the threshold (e.g., the strident /z/) and others (e.g., /v/) are below. This problem is also language-dependent; e.g., English appears to have relatively weak voice bars, while French has strong ones.

Unlike short-time energy, the ZCR is highly sensitive to noise in the recording environment (e.g., 60 Hz hum from a power supply) or in analog-to-digital (A/D) conversion. Since energy below 100 Hz is largely irrelevant for speech processing, it may be desirable to highpass filter the speech in addition to the normal lowpass filtering before A/D conversion.

The ZCR can be applied to speech recognition. If speech is first passed through a bank of bandpass filters, each filter's output better resembles a narrowband signal, whose frequency of major energy concentration the ZCR easily estimates. Such a frequency could be a single harmonic (for filter bandwidths less than F0) or a formant frequency (for bandwidths of about 300–500 Hz). A bank or eight filters covering the 0–4 kHz range provides a simple set of eight measures, which could replace a more complex spectral representation (e.g., a DFT) in some applications.

6.3.4 Short-Time Aurocorrelation Function

The Fourier transform $S(e^{j\omega})$ of speech $s(n)$ provides both spectral magnitude and phase. The time signal $r(k)$ for the inverse Fourier transform of the energy spectrum ($|S(e^{j\omega})|^2$) is called the *autocorrelation* of $s(n)$. $r(k)$ preserves information about harmonic and formant amplitudes in $s(n)$ as well as its periodicity, while ignoring phase (as do many applications), since phase is less important perceptually and carries much less communication information than spectral magnitude. $r(k)$ has applications in F0 estimation, voiced/unvoiced determination, and linear prediction.

The autocorrelation function is a special case of the cross-correlation function,

$$\phi_{sy}(k) = \sum_{m=-\infty}^{\infty} s(m)y(m-k), \tag{6.9}$$

which measures the similarity of two signals $s(n)$ and $y(n)$ as a function of the time delay between them. By summing the products of a signal sample and a delayed sample from another signal, the cross-correlation is large if at some delay the two signals have similar waveforms. The range of summation is usually limited (i.e., windowed), and the function can be normalized by dividing by the number of summed samples.

When the same signal is used for $s(n)$ and $y(n)$, Equation (6.9) yields an autocorrelation. It is an even function ($r(k) = r(-k)$), it has maximum value at $k = 0$, and $r(0)$ equals the energy in $s(n)$ (or average power, for random or periodic signals). If $s(n)$ is periodic in P samples, then $r(k)$ also has period P. Maxima in $r(k)$ occur for $k = 0, \pm P, \pm 2P$, etc., independently of the absolute timing of the pitch periods; i.e., the window does not have to be placed synchronously with the pitch periods.

The *short-time* autocorrrelation function is obtained by windowing $s(n)$ and then using Equation (6.9), yielding

$$R_n(k) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)s(m-k)w(n-m+k). \tag{6.10}$$

Equivalently, the product of speech $s(n)$ with its delayed version $s(n-k)$ is passed through a filter with response $w(n)w(n+k)$ (time index n indicates the position of the window). Equation (6.10) is evaluated for different values of k depending on the application. For linear prediction (Section 6.5), $R_n(k)$ for k ranging from 0 to 10–16 are typically needed, depending on the signal bandwidth. In F0 determination, $R_n(k)$ is needed for k near the estimated number of samples in a pitch period; if no suitable prior F0 estimate is available, $R_n(k)$ is calculated for k from the shortest possible period (perhaps 3 ms for a female voice) to the longest (e.g., 20 ms for men). With a sampling rate of 10,000 samples/s, the latter approach can require up to 170 calculations of $R_n(k)$ for each speech frame, if a pitch period resolution of 0.1 ms is desired.

Short windows minimize calculation: if $w(n)$ has N samples, $N-k$ products are needed for each value of $R_n(k)$. Proper choice of $w(n)$ also helps; e.g., using a rectangular window reduces the number of multiplications; symmetries in autocorrelation calculation can also be exploited (see LPC below). While the duration of $w(n)$ is almost directly proportional to the calculation (especially if $N \gg k$), there is a conflict between minimizing N to save computation and having enough speech samples in the window to yield a valid autocorrelation function: longer $w(n)$ give better frequency resolution. For F0 estimation, $w(n)$ must include more than one pitch period, so that $R_n(k)$ exhibits periodicity and the corresponding energy spectrum $|X_n(e^{j\omega})|^2$ resolves individual harmonics of F0 (see Figure 6.4). Spectral estimation applications (e.g., LPC) permit short windows since harmonic resolution is unimportant and the formant spectrum can be found from a portion of a pitch period.

For F0 estimation, an alternative to using autocorrelation is the average magnitude difference function (AMDF) [4]. Instead of multiplying speech $s(m)$ by $s(m-k)$, the

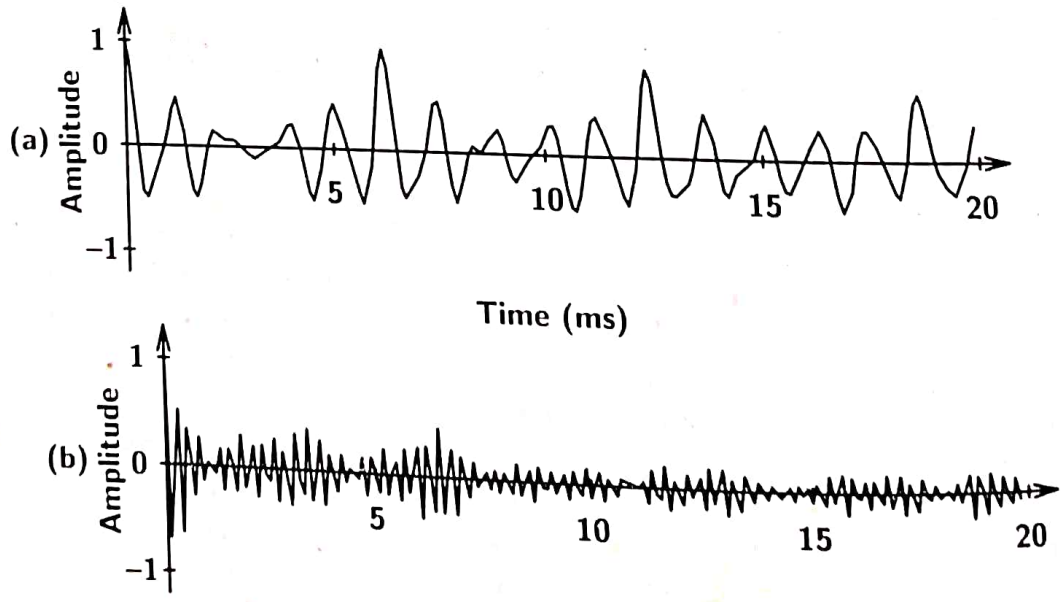

Figure 6.7 Typical autocorrelation function for (a) voiced speech and (b) unvoiced speech, using a 20 ms rectangular window ($N = 201$).

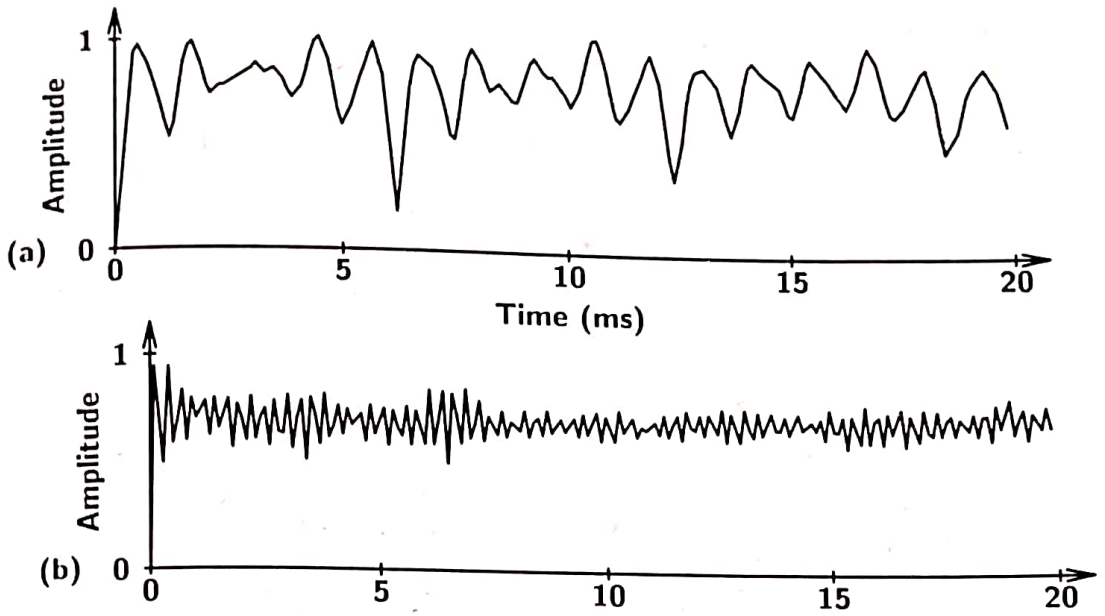

Figure 6.8 AMDF function (normalized to 1.0) for the same speech segments as in Figure 6.7.

magnitude of their difference is taken:

$$\text{AMDF}(k) = \sum_{m=-\infty}^{\infty} |s(m) - s(m-k)|. \tag{6.11}$$

Since subtraction and rectification are much simpler operations than multiplication, the AMDF is considerably faster. Where $R_n(k)$ peaks for values of k near multiples of the pitch period (Figure 6.7), the AMDF has minima (Figure 6.8).

Some speech recognition applications have used a simplified version of the autocorrelation [5]:

$$\psi(k) = \sum_{m=-\infty}^{\infty} \text{sgn}(s(m))s(m-k). \tag{6.12}$$

Replacing $s(m)$ by its sign in Equation (6.9) eliminates the need for multiplications and reduces the emphasis that $r(k)$ normally places on the high-amplitude portions of $s(n)$.

6.4 FREQUENCY-DOMAIN (SPECTRAL) PARAMETERS

The frequency domain provides most useful parameters for speech processing. Speech signals are more consistently and easily analyzed spectrally than in the time domain. The basic model of speech production with a noisy or periodic waveform that excites a vocal tract filter corresponds well to separate spectral models for the excitation and for the vocal tract. Repeated utterances of a sentence by a speaker often differ greatly temporally while being very similar spectrally. Human hearing appears to pay much more attention to spectral aspects of speech (e.g., amplitude distribution in frequency) than to phase or timing aspects. Thus, spectral analysis is used to extract most parameters from speech.