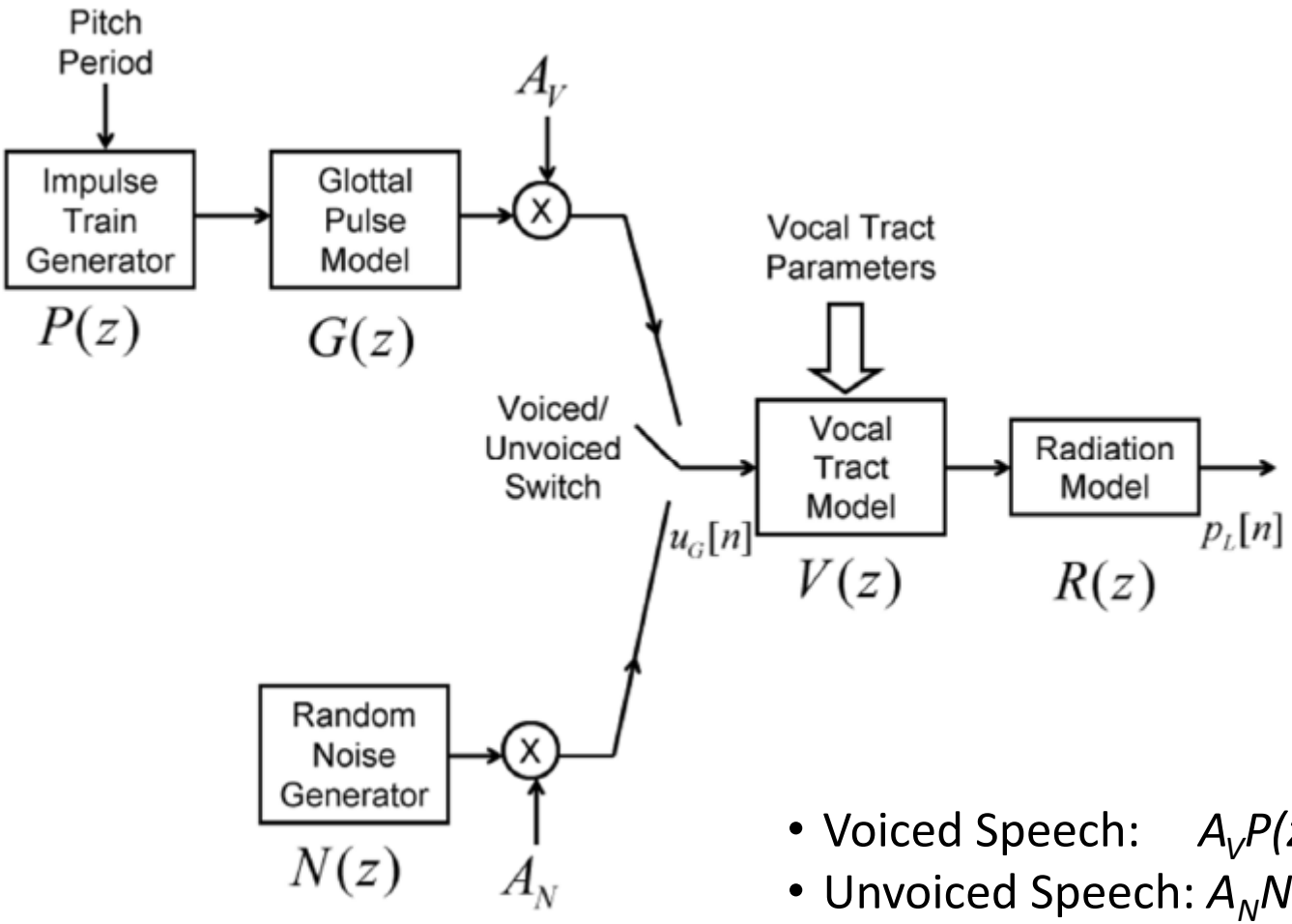
Chapter 7

Frequency-Domain Representations
语音信号的频域表征

General Discrete-Time Model of Speech Production



- Voiced Speech: $A_V P(z)G(z)V(z)R(z)$
- Unvoiced Speech: $A_N N(z)V(z)R(z)$

DTFT and DFT of Speech

- The DTFT and the DFT for the speech signal could be calculated by the following:

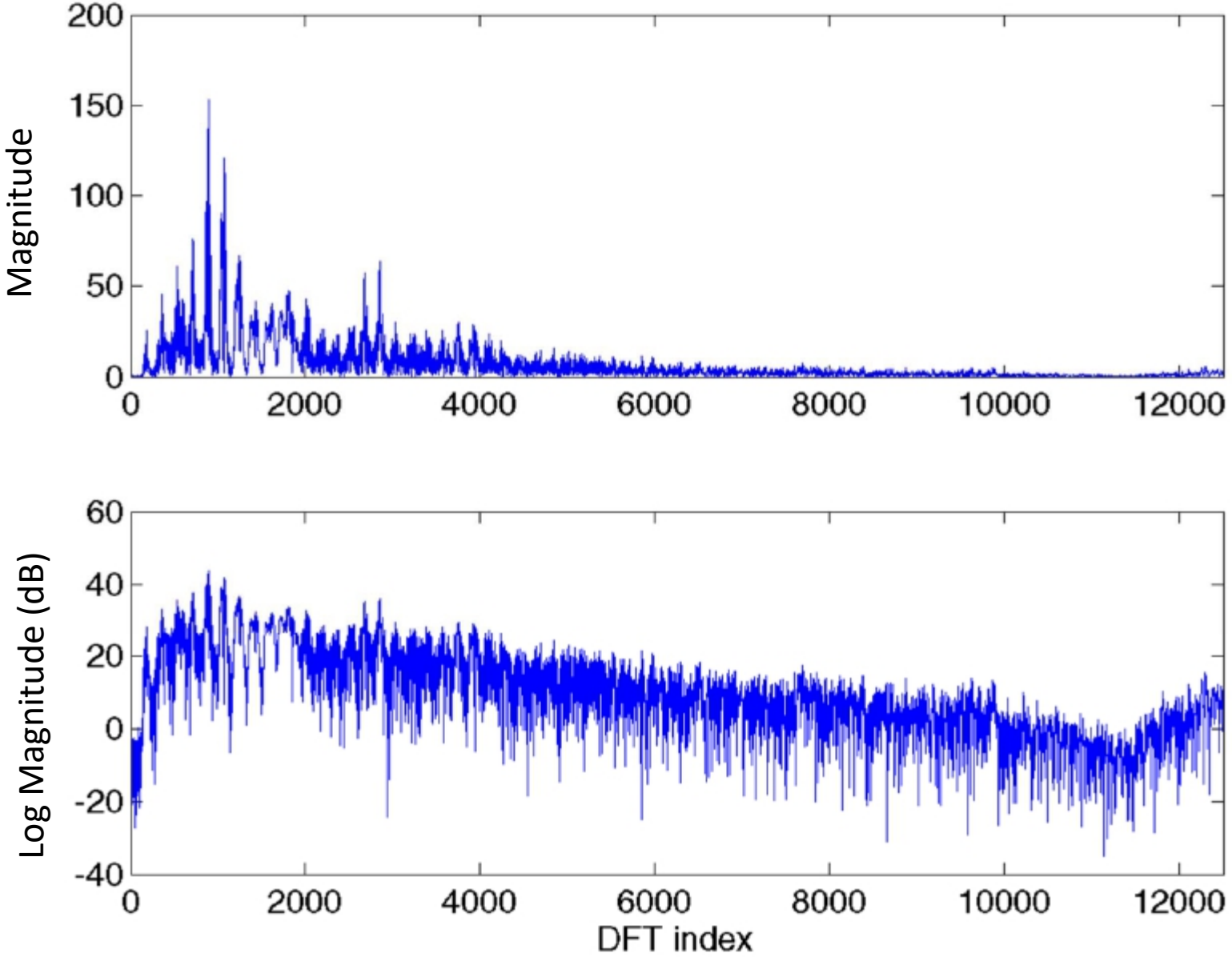
$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m} \quad (DTFT)$$

$$X(k) = \sum_{m=0}^{L-1} x(m)\, e^{-j(2\pi/L)km}, \quad k = 0,1,\ldots,L-1$$

$$= X(e^{j\omega})\Big|_{\omega=(2\pi k/L)} \qquad (DFT)$$

using a value of L=25000 we get the following plot

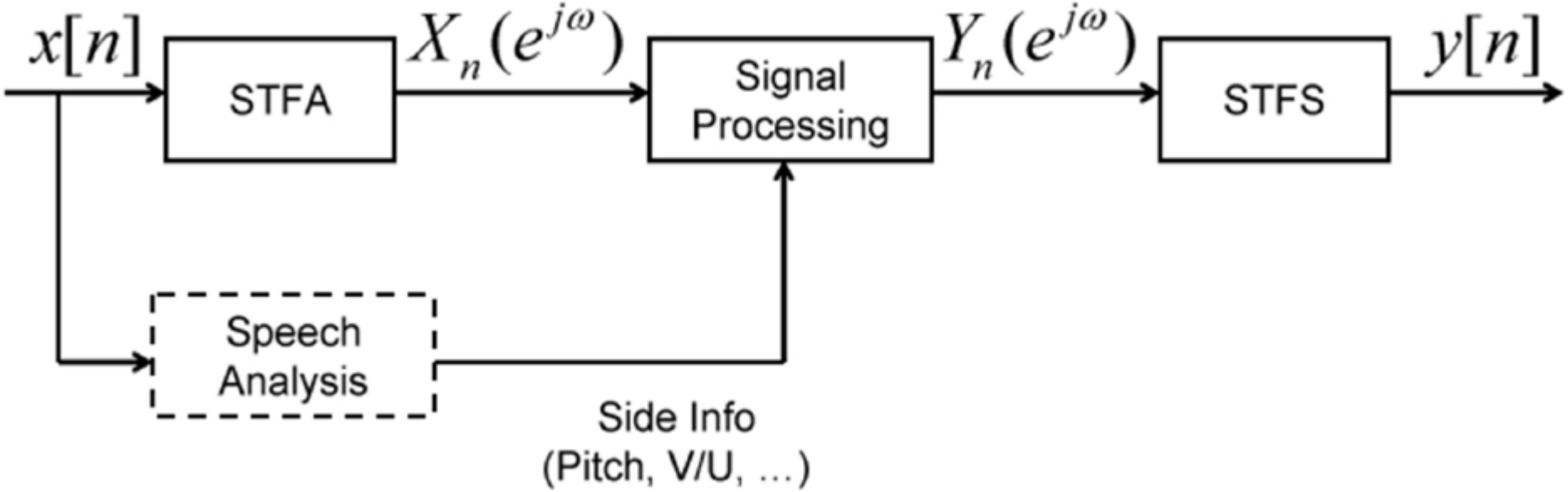
25000-Point DFT of Speech



Why STFT for Speech Signals

- steady state sounds, like vowels, are produced by periodic excitation of a linear system => speech spectrum is the product of the excitation spectrum and the vocal tract frequency response
- speech is a time-varying signal => need more sophisticated analysis to reflect time varying properties
  - changes occur at syllabic rates (~10 times/sec)
  - over fixed time intervals of 10-30 msec, properties of most speech signals are relatively constant

Frequency Domain Processing



- Coding
  - transform, subband, homomorphic, channel vocoders
- Restoration/Enhancement/Modification
  - noise and reverberation removal, time-scale modifications (speed-up and slow-down of speech)

Overview of Lecture

- define time-varying Fourier transform (STFT) analysis method
- define synthesis method from time-varying FT (filter-bank summation, overlap addition)
- show how time-varying FT can be viewed in terms of a bank of filters model
- computation methods based on using FFT
- application to vocoders, spectrum displays, format estimation, pitch period estimation

Short-Time Fourier Transform (STFT)

Short-Time Fourier Transform

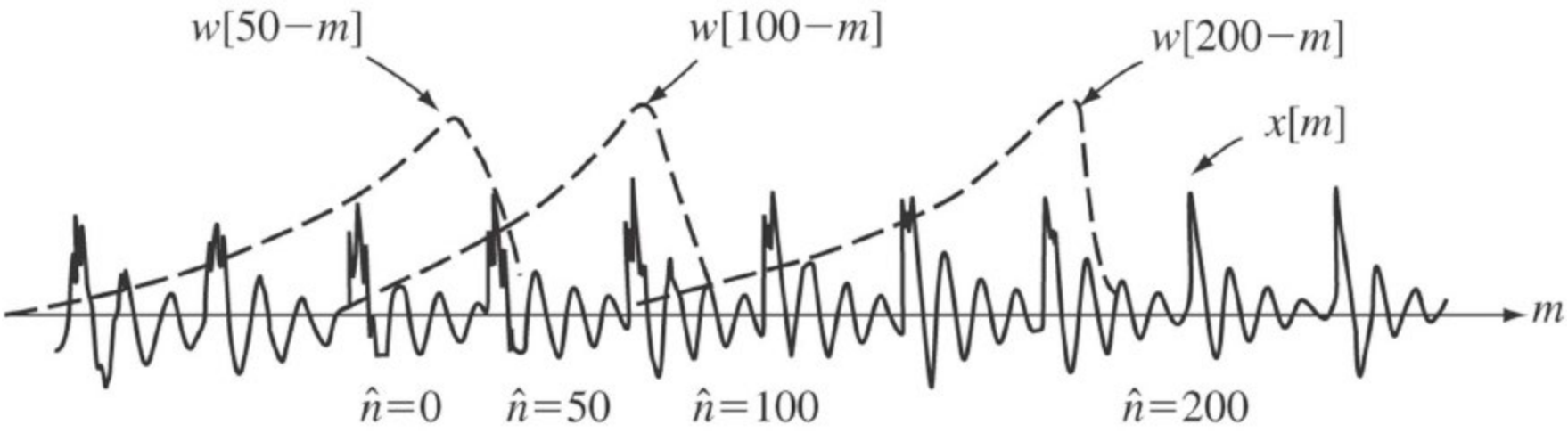
- speech is not a stationary signal, i.e., it has properties that change with time
- thus a single representation based on all the samples of a speech utterance, for the most part, has no meaning
- instead, we define a time-dependent Fourier transform (TDFT or STFT) of speech that changes periodically as the speech properties change over time

Definition of STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

both $\hat{n}$ and $\hat{\omega}$ are variables

- $w(\hat{n}-m)$ is a real window which determines the portion of $x(\hat{n})$ that is used in the computation of $X_{\hat{n}}(e^{j\hat{\omega}})$

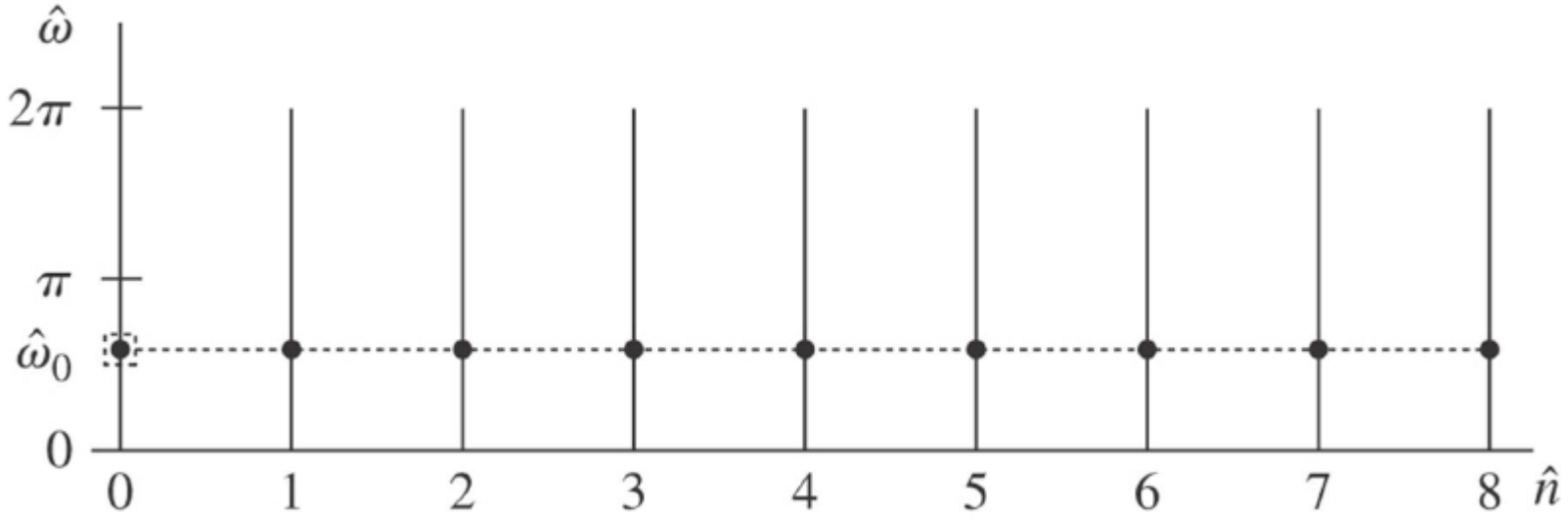


Short-Time Fourier Transform

- STFT is a function of two variables, the time index, $\hat{n}$, which is discrete, and the frequency variable, $\hat{\omega}$, which is continuous

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

$$= DTFT\,(x(m)w(\hat{n}-m)) \Rightarrow \hat{n} \text{ fixed, } \hat{\omega} \text{ variable}$$



STFT-Different Time Origins

- the STFT can be viewed as having two different time origins
  1. time origin tied to signal $x(n)$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$
$$= DTFT\left[x(m)w(\hat{n}-m)\right], \quad \hat{n} \text{ fixed, } \hat{\omega} \text{ variable}$$

  2. time origin tied to window signal $w(-m)$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = e^{-j\hat{\omega}\hat{n}} \sum_{m=-\infty}^{\infty} x(\hat{n}+m)w(-m)e^{-j\hat{\omega}m}$$
$$= e^{-j\hat{\omega}\hat{n}}\tilde{X}(e^{j\hat{\omega}})$$
$$= e^{-j\hat{\omega}\hat{n}}\ DTFT\left[w(-m)x(\hat{n}+m)\right], \quad \hat{n} \text{ fixed, } \hat{\omega} \text{ variable}$$

Interpretations of STFT

- there are 2 distinct interpretations of $X_{\hat{n}}(e^{j\hat{\omega}})$
  1. assume $\hat{n}$ is fixed, then $X_{\hat{n}}(e^{j\hat{\omega}})$ is simply the normal Fourier transform of the sequence $w(\hat{n}-m)x(m), -\infty < m < \infty$

     => for fixed $\hat{n}$, $X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform
  2. consider $X_{\hat{n}}(e^{j\hat{\omega}})$ as a function of the time index $\hat{n}$ with $\hat{\omega}$ fixed. Then $X_{\hat{n}}(e^{j\hat{\omega}})$ is in the form of a convolution of the signal $x(\hat{n})e^{-j\hat{\omega}\hat{n}}$ with the window $w(\hat{n})$. This leads to an interpretation in the form of linear filtering of the frequency modulated signal $x(\hat{n})e^{-j\hat{\omega}\hat{n}}$ by $w(\hat{n})$.
- We will now consider each of these interpretations of the STFT in a lot more detail

DTFT Interpretation of STFT

Fourier Transform Interpretation

- consider $X_{\hat{n}}(e^{j\hat{\omega}})$ as the normal Fourier transform of the sequence $w(\hat{n}-m)x(m), -\infty < m < \infty$ for fixed $\hat{n}$
- the window $w(\hat{n}-m)$ slides along the sequence $x(m)$ and defines a new STFT for every value of $\hat{n}$
- what are the conditions for the existence of the STFT
  - the sequence $w(\hat{n}-m)x(m)$ must be absolutely summable for all values of $\hat{n}$
    - since $\left|x(\hat{n})\right| \le L$ (32767 for 16-bit sampling)
    - since $\left|w(\hat{n})\right| \le 1$ (normalized window level)
    - since window duration is usually finite
  - $w(\hat{n}-m)x(m)$ is absolutely summable for all $\hat{n}$

Signal Recovery from STFT

- since for a given value of $\hat{n}$, $X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform, we can recover the input sequence exactly
- since $X_{\hat{n}}(e^{j\hat{\omega}})$ is the normal Fourier transform of the window sequence $w(\hat{n}-m)x(m)$, then

$$w(\hat{n}-m)\,x(m)=\frac{1}{2\pi}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}$$

- assuming the window satisfies the property that $w(0)\neq 0$ a trivial requirement), then by evaluating the inverse Fourier transform when $m=\hat{n}$, we obtain

$$x(\hat{n})=\frac{1}{2\pi w(0)}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

Signal Recovery from STFT

$$x(\hat{n}) = \frac{1}{2\pi w(0)} \int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}}) e^{j\omega\hat{n}} d\hat{\omega}$$

- with the requirement that $w(0) \neq 0$, the sequence $x(\hat{n})$ can be recovered from $X_{\hat{n}}(e^{j\hat{\omega}})$, if $X_{\hat{n}}(e^{j\hat{\omega}})$ is known for all values of $\hat{\omega}$ over one complete period
  - sample-by-sample recovery process
  - $X_{\hat{n}}(e^{j\hat{\omega}})$ must be known for every value of $\hat{n}$ and for all $\hat{\omega}$
- can also recover sequence $w(\hat{n}-m)x(m)$ but can't guarantee that $x(m)$ can be recovered since $w(\hat{n}-m)$ can equal 0

Alternative Forms of STFT

1. real and imaginary parts

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \mathrm{Re}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right] + j\,\mathrm{Im}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right]$$
$$= a_{\hat{n}}(\hat{\omega}) - j\,b_{\hat{n}}(\hat{\omega})$$
$$a_{\hat{n}}(\hat{\omega}) = \mathrm{Re}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right]$$
$$b_{\hat{n}}(\hat{\omega}) = -\mathrm{Im}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right]$$

- when $x(m)$ and $w(\hat{n}-m)$ are both real (usually the case) can show that $a_{\hat{n}}(\hat{\omega})$ is symmetric in $\hat{\omega}$, and $b_{\hat{n}}(\hat{\omega})$ is anti-symmetric in $\hat{\omega}$

2. magnitude and phase

$$X_{\hat{n}}(e^{j\hat{\omega}}) = |X_{\hat{n}}(e^{j\hat{\omega}})|e^{j\theta_{\hat{n}}(\hat{\omega})}$$

- can relate $|X_{\hat{n}}(e^{j\hat{\omega}})|$ and $\theta_{\hat{n}}(\hat{\omega})$ to $a_{\hat{n}}(\hat{\omega})$ and $b_{\hat{n}}(\hat{\omega})$

Role of Window in STFT

- The window $w(\hat{n}-m)$ does the following
  - chooses portion of $x(m)$ to be analyzed
  - window shape determines the nature of $X_{\hat{n}}(e^{j\hat{\omega}})$
- Since $X_{\hat{n}}(e^{j\hat{\omega}})$ (for fixed $\hat{n}$) is the normal FT of $w(\hat{n}-m)x(m)$ then if we consider the normal FT's of both $x(n)$ and $w(n)$ individually, we get

$$X(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}$$

$$W(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} w(m)e^{-j\hat{\omega}m}$$

Role of Window in STFT

- then for fixed $\hat{n}$, the normal Fourier transform of the product $w(\hat{n}-m)x(m)$ is the convolution of the transforms of $w(\hat{n}-m)$ and $x(m)$

- for fixed $\hat{n}$, the FT of $w(\hat{n}-m)$ is $W(e^{-j\hat{\omega}})e^{-j\hat{\omega}\hat{n}}$ --thus

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\theta})e^{-j\theta\hat{n}}\, X(e^{j(\hat{\omega}-\theta)})d\theta$$

- limiting case

$$w(\hat{n}) = 1 \quad -\infty < \hat{n} < \infty \Leftrightarrow W(e^{j\hat{\omega}}) = 2\pi\delta(\hat{\omega})$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} 2\pi\delta(-\theta)\, X(e^{j(\hat{\omega}-\theta)})e^{-j\theta\hat{n}}d\theta = X(e^{j\hat{\omega}})$$

  - we get the same thing no matter where the window is shifted

Interpretation of Role of Window

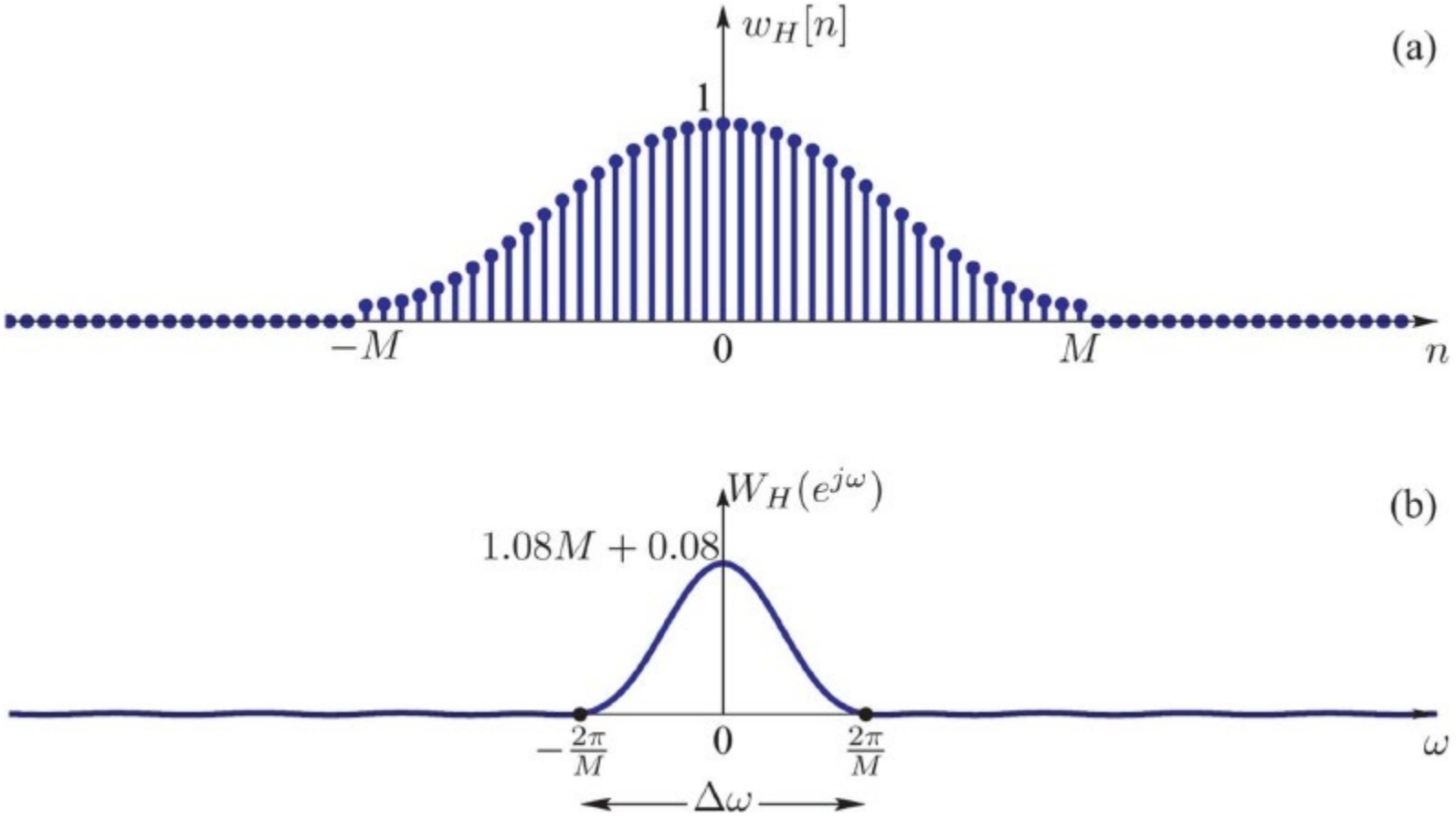
- $X_{\hat{n}}(e^{j\hat{\omega}})$ is the convolution of $X(e^{j\hat{\omega}})$ with the FT of the shifted window sequence $W(e^{-j\hat{\omega}})e^{-j\hat{\omega}\hat{n}}$
- $X(e^{j\hat{\omega}})$ really doesn't have meaning since $x(\hat{n})$ varies with time
- consider $x(\hat{n})$ defined for window duration and extended for all time to have the same properties

  => then $X(e^{j\hat{\omega}})$ does exist with properties that reflect the sound within the window
- $X_{\hat{n}}(e^{j\hat{\omega}})$ is a smoothed version of the FT of the part of $x(\hat{n})$ that is within the window w

Windows in STFT

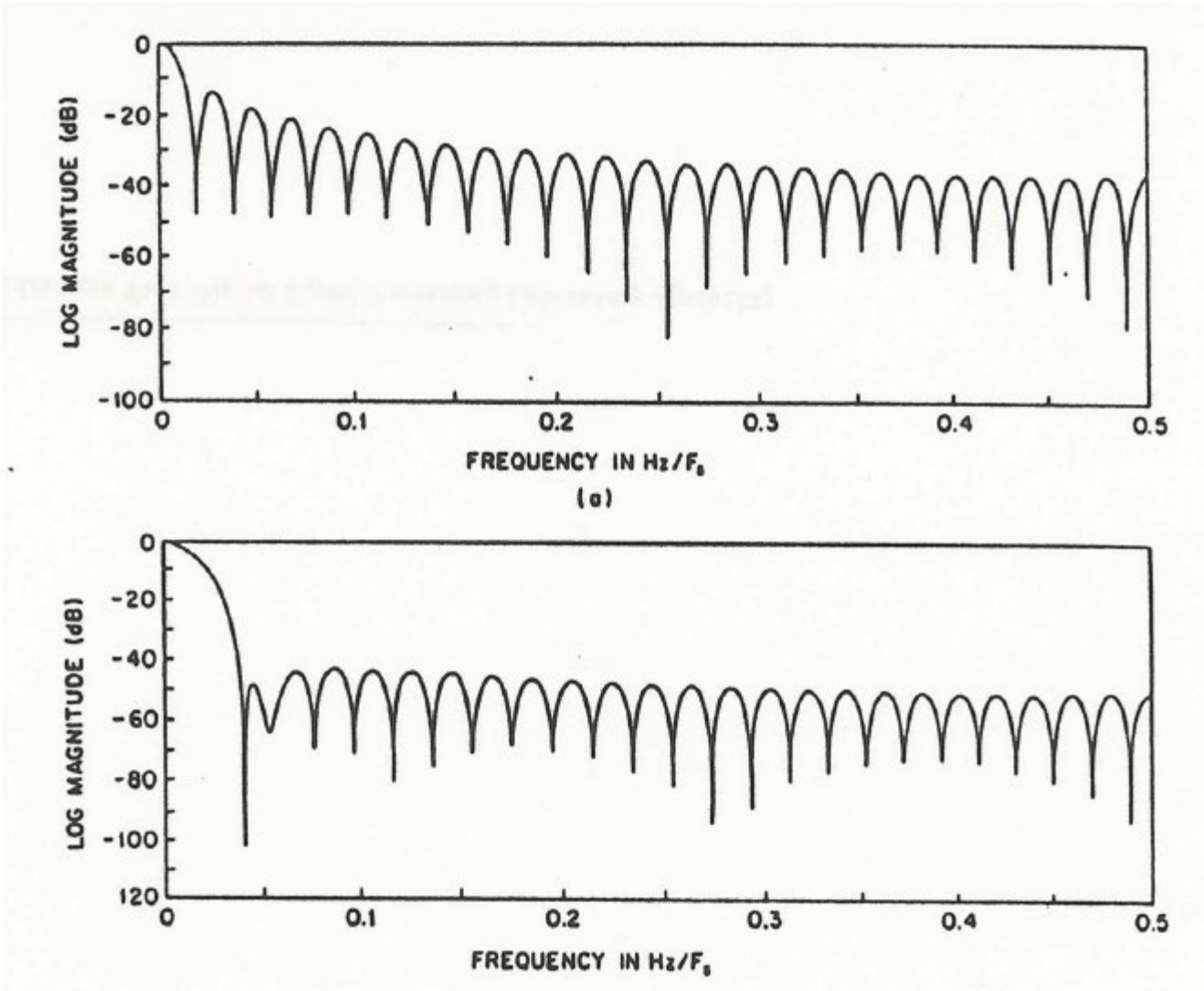
- consider rectangular and Hamming windows, where width of the main spectral lobe is inversely proportional to window length, and side lobe levels are essentially independent of window length
  - Rectangular Window: flat window of length L samples; first zero in frequency response occurs at F_S/L, with sidelobe levels of -14 dB or lower
  - Hamming Window: raised cosine window of length L samples; first zero in frequency response occurs at $2\,F_S/L$, with sidelobe levels of -40 dB or lower

Windows

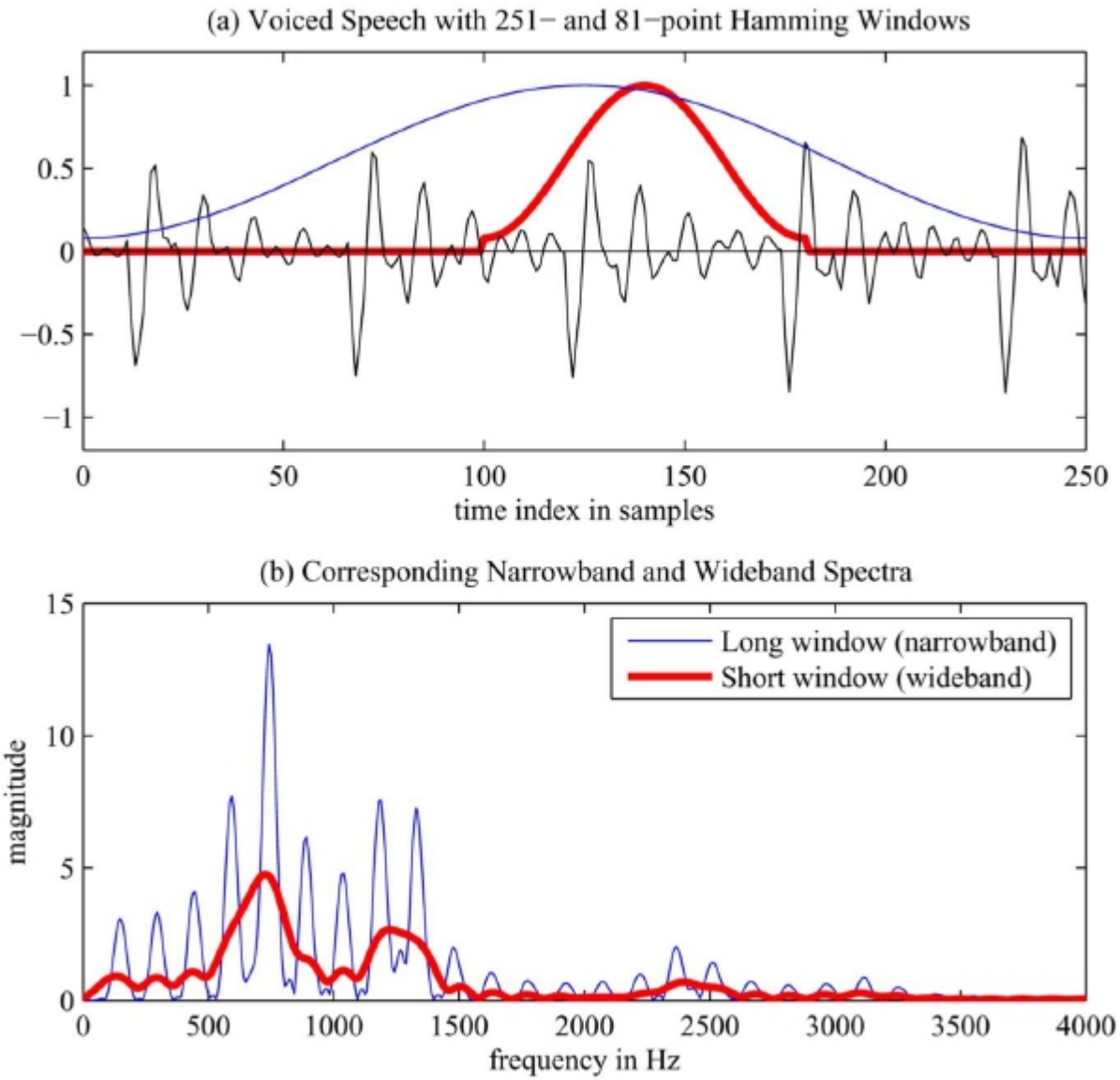


$L=2M+1$-point Hamming window and its corresponding DTFT

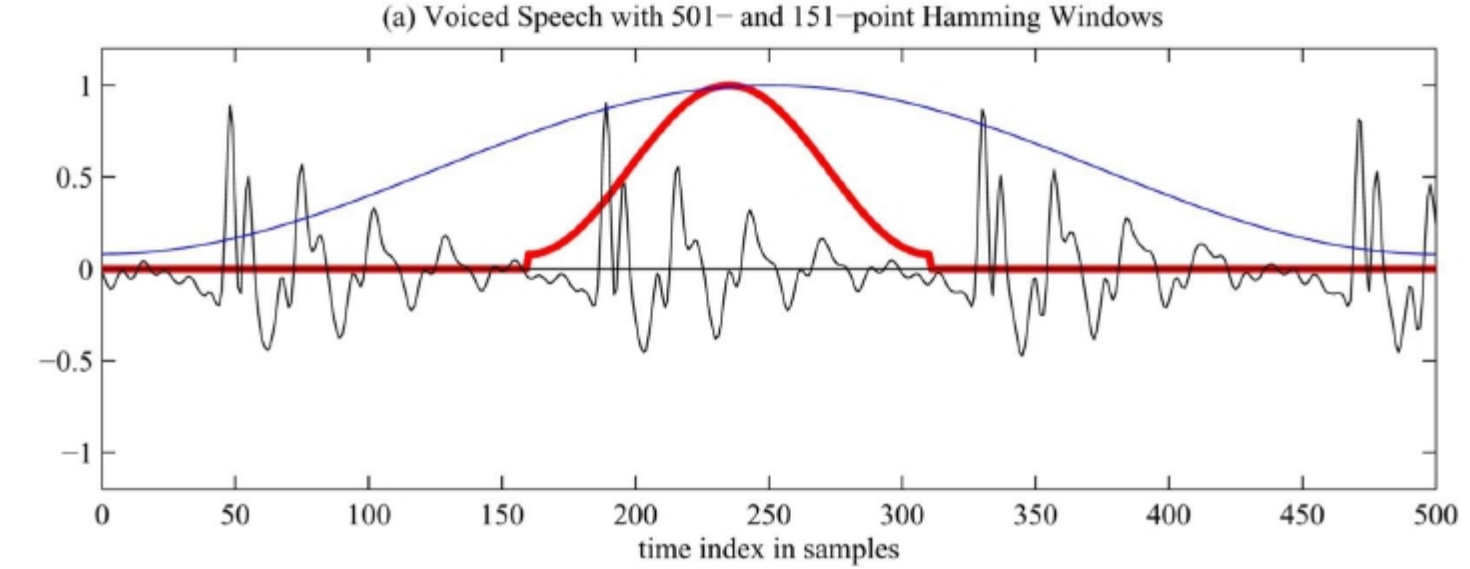
Frequency Responses of Windows

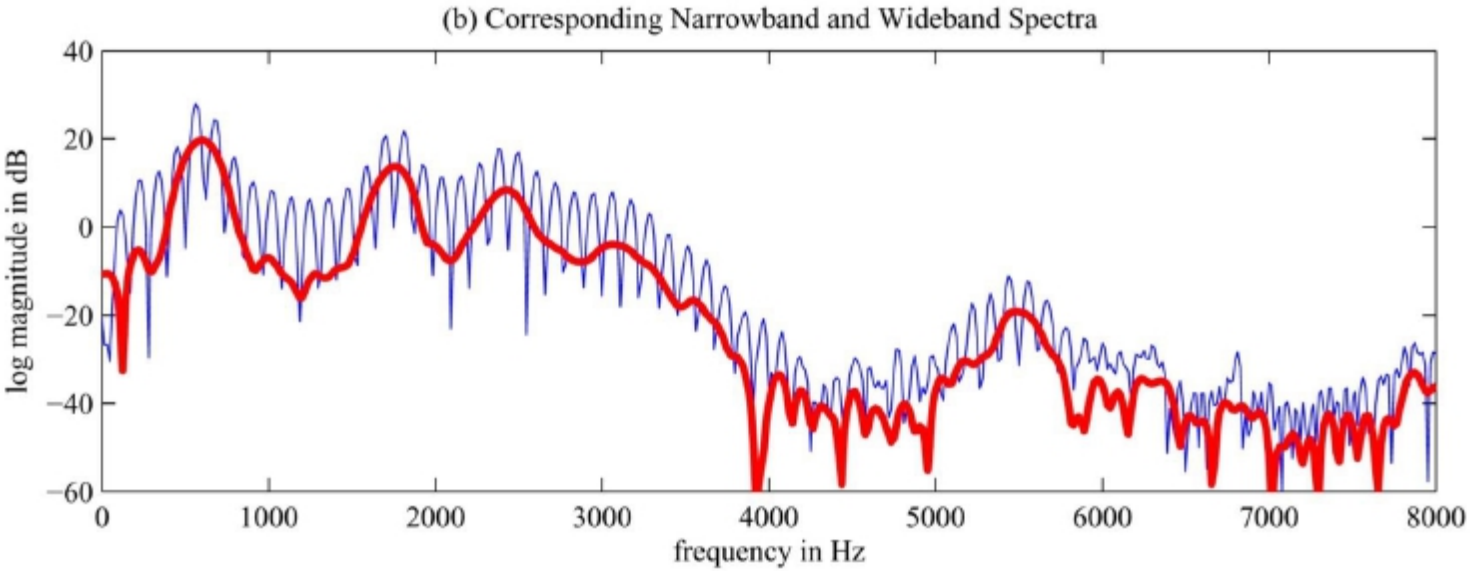


Effect of Window Length - HW

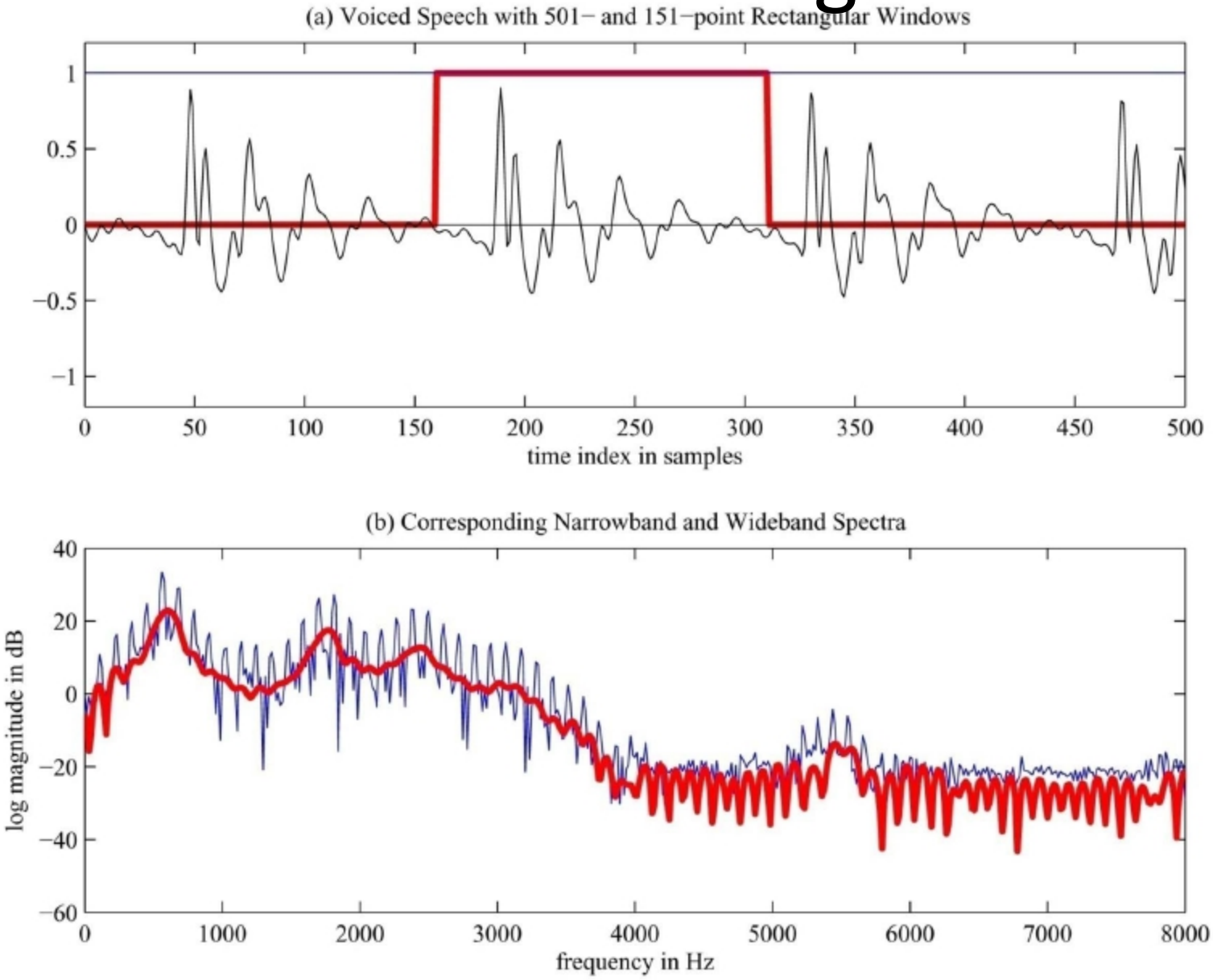


Effect of Window Length - HW

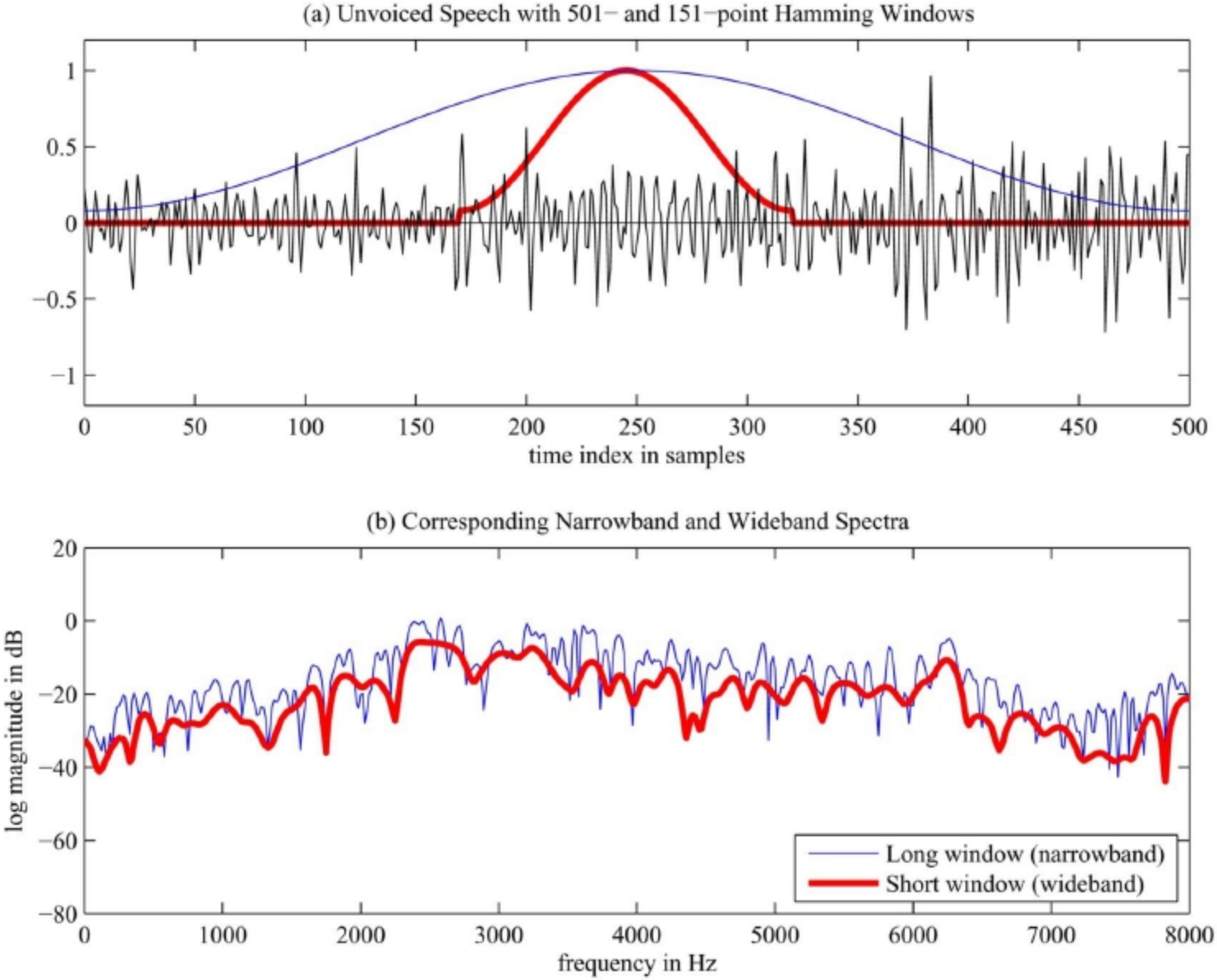




Effect of Window Length - RW



Effect of Window Length - HW



Relation to Short-Time Autocorrelation

- $X_{\hat{n}}(e^{j\hat{\omega}})$ is the discrete-time Fourier transform of $w[\hat{n}-m]x[m]$ for each value of $\hat{n}$, then it is seen that

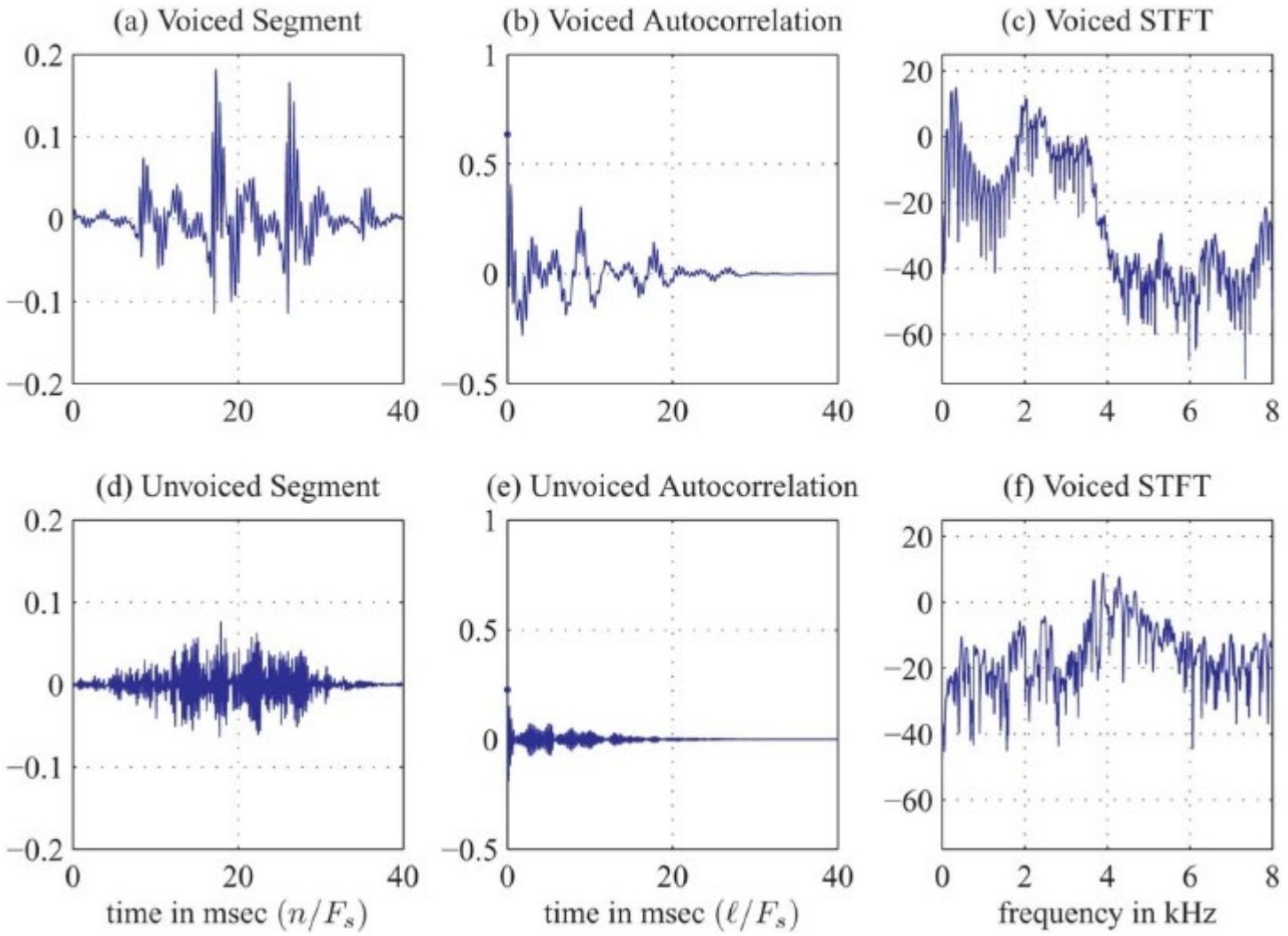
$$S_{\hat{n}}(e^{j\hat{\omega}}) = |X_{\hat{n}}(e^{j\hat{\omega}})|^2 = X_{\hat{n}}(e^{j\hat{\omega}})X_{\hat{n}}^*(e^{j\hat{\omega}})$$

  is the Fourier transform of

$$R_{\hat{n}}(l) = \sum_{m=-\infty}^{\infty} w[\hat{n}-m]x[m]w[\hat{n}-l-m]x[m+l]$$

  which is the short-time autocorrelation function of the previous chapter. Thus the above equations relate the short-time spectrum to the short-time autocorrelation.

Short-Time Autocorrelation and STFT



Summary of FT view of STFT

- Interpret $X_{\hat{n}}(e^{j\omega})$ as the normal Fourier transform of the sequence $w(\hat{n}-m)x(m), -\infty < m < \infty$
- properties of this Fourier transform depend on the window
  - frequency resolution of $X_{\hat{n}}(e^{j\omega})$ varies inversely with the length of the window => want long windows for high resolution
  - want x(n) to be relatively stationary (non-time-varying) during duration of window for most stable spectrum => want short windows

- as usual in speech processing, there needs to be a compromise between good temporal resolution (short windows) and good frequency resolution (long windows)

Linear Filtering Interpretation of STFT

Linear Filtering Interpretation

1. modulation-lowpass filter form

$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}w(n-m)$$

$$= w(n) * \left(x(n)e^{-j\hat{\omega}n}\right), \quad n \text{ variable, } \hat{\omega} \text{ fixed}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})X(e^{j(\theta+\hat{\omega})})e^{j\theta n}d\theta$$

2. bandpass filter-demodulation

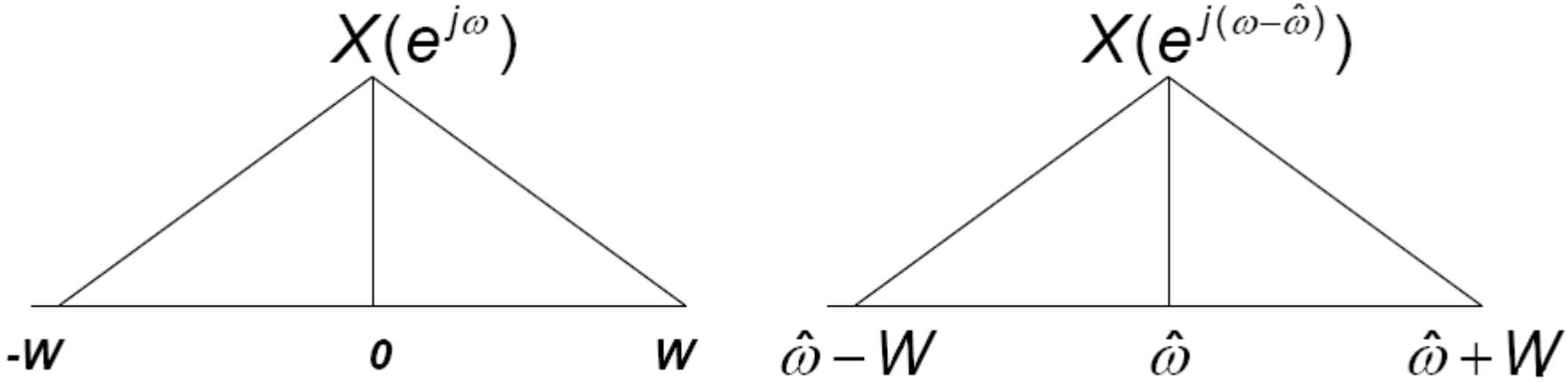
$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} w(m)x(n-m)e^{-j\hat{\omega}(n-m)}$$

$$= e^{-j\hat{\omega}n}\sum_{m=-\infty}^{\infty}(w(m)e^{j\hat{\omega}m})x(n-m)$$

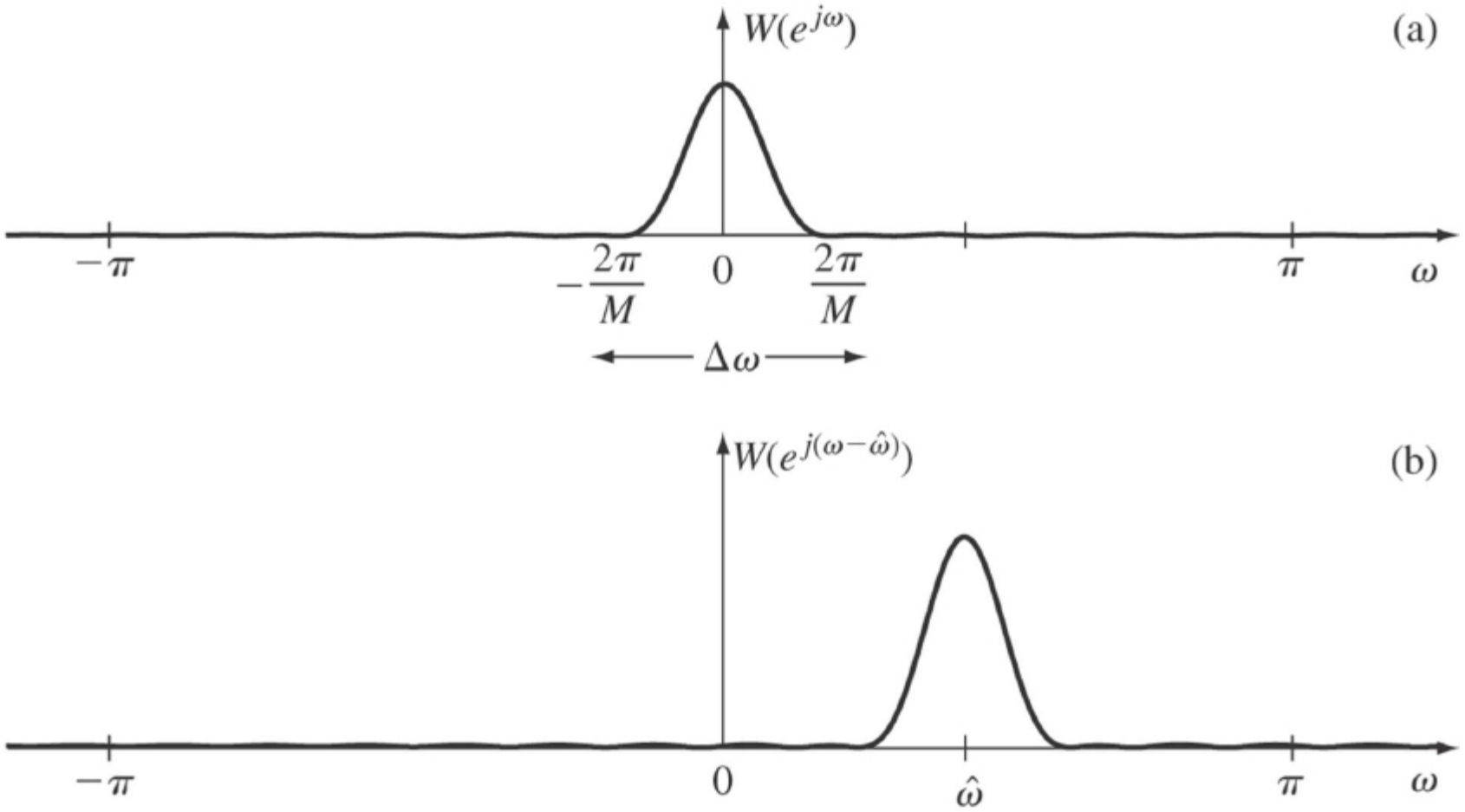
$$= e^{-j\hat{\omega}n}[(w(n)e^{j\hat{\omega}n}) * x(n)], \quad n \text{ variable, } \hat{\omega} \text{ fixed}$$

Linear Filtering Interpretation

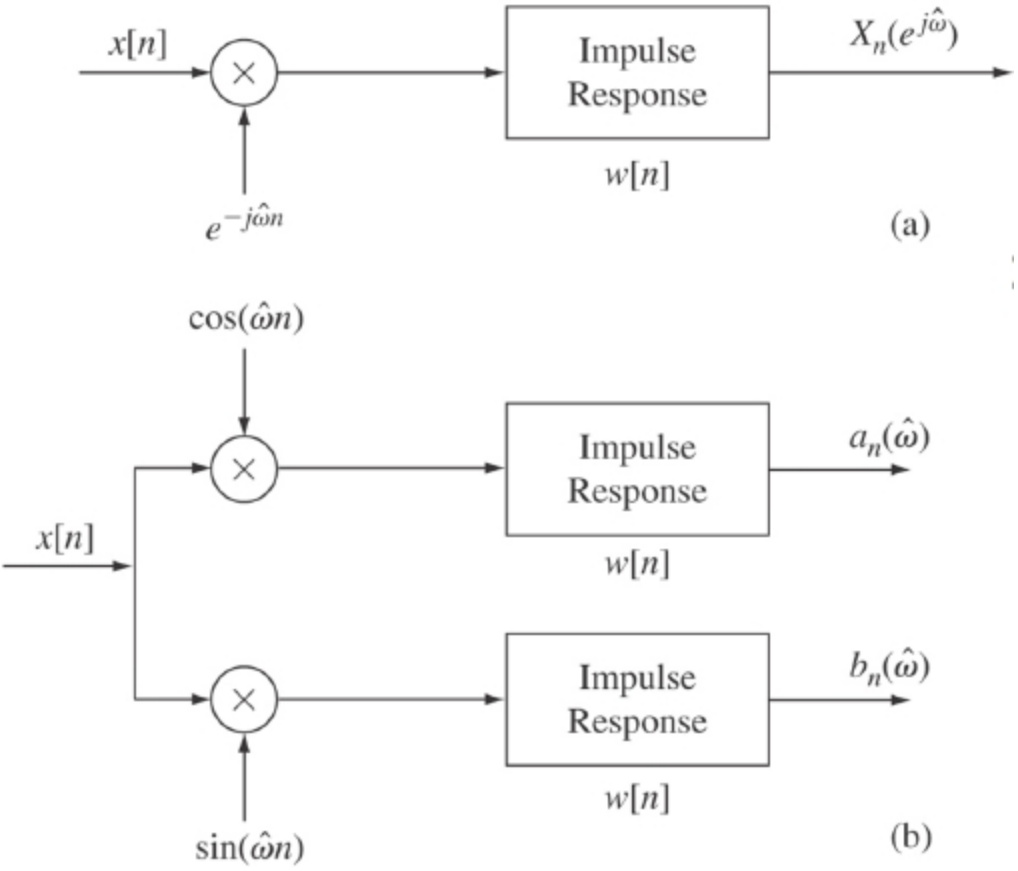
$$x[n]e^{j\hat{\omega}n} \leftrightarrow X(e^{j\omega}) * FT(e^{j\hat{\omega}n})$$
$$= X(e^{j\omega}) * \delta(\omega - \hat{\omega})$$
$$= X(e^{j(\omega-\hat{\omega})})$$



Linear Filtering Interpretation



Linear Filtering Interpretation



1. modulation-lowpass filter form:

$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}w(n-m),$$

n variable, $\hat{\omega}$ fixed

$$= \left(x(n)e^{-j\hat{\omega}n}\right) * w(n)$$

$$= \left(x(n)\cos(\hat{\omega}n)\right) * w(n) -$$

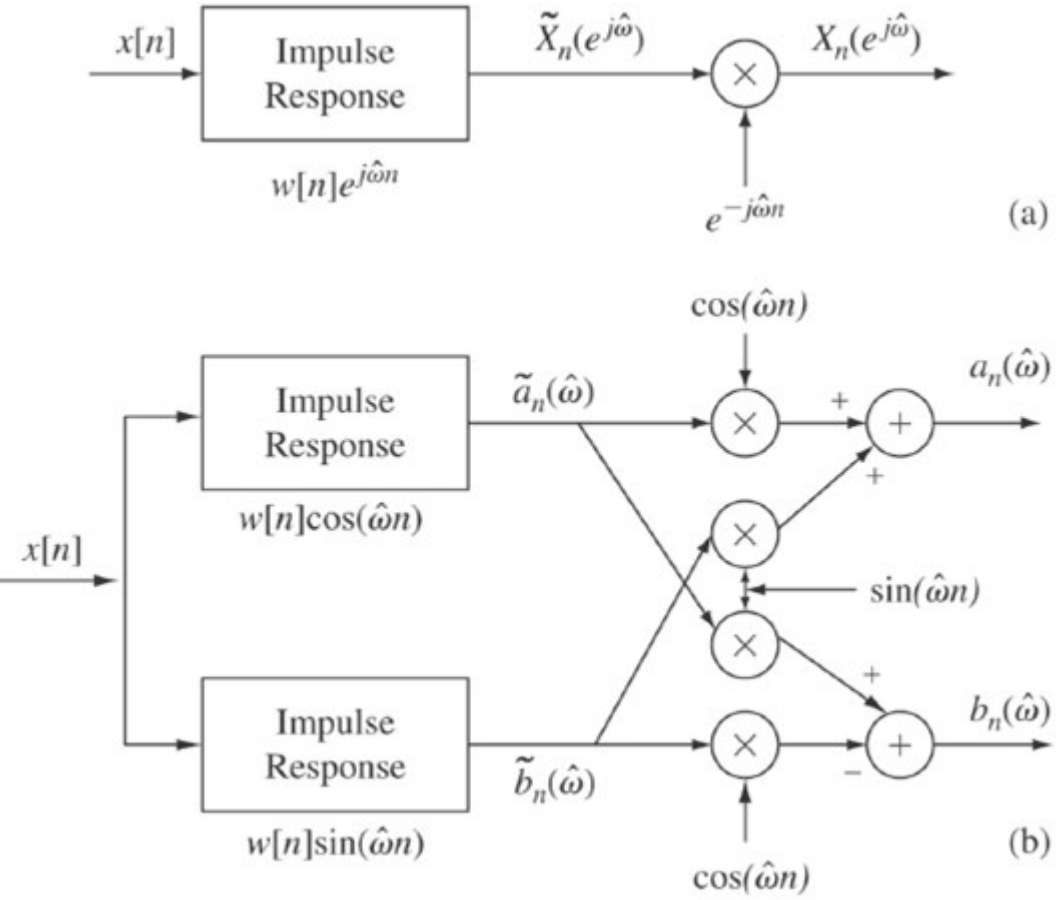
$$j\left(x(n)\sin(\hat{\omega}n)\right) * w(n)$$

$$= a_n(\hat{\omega}) - jb_n(\hat{\omega})$$

Linear Filtering Interpretation

2. bandpass filter-demodulation form

$$X_n(e^{j\hat{\omega}}) = e^{-j\hat{\omega}n}\left[\left(w(n)e^{j\hat{\omega}n}\right) * x(n)\right],\ n \text{ variable},\ \hat{\omega} \text{ fixed}$$



- complex bandpass filter output modulated by signal $e^{-j\hat{\omega}n}$
- if $W(e^{j\theta})$ is lowpass, then filter is bandpass around $\theta = \hat{\omega}$

Summary - STFT

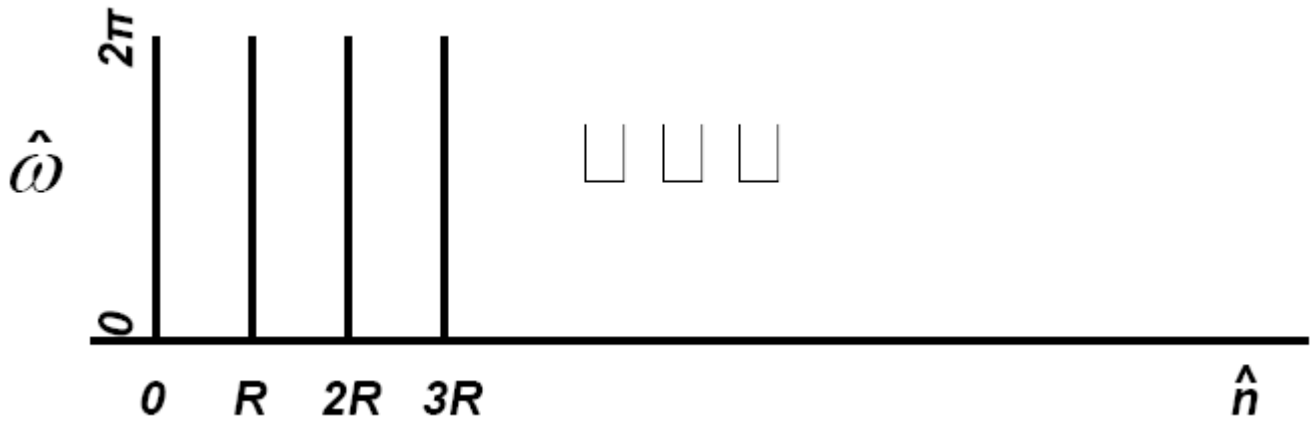
$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m},$$

$$-\infty < \hat{n} < \infty, 0 \le \hat{\omega} < 2\pi$$

- Fixed value of $\hat{n}$, varying $\hat{\omega}$ -- DFT Interpretation
- Fixed value of $\hat{\omega}$, varying $\hat{n}$ -- Filter Bank Interpretation

Summary – DFT Interpretation

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m},\ -\infty < \hat{n} < \infty,\ 0 \le \hat{\omega} < 2\pi$$



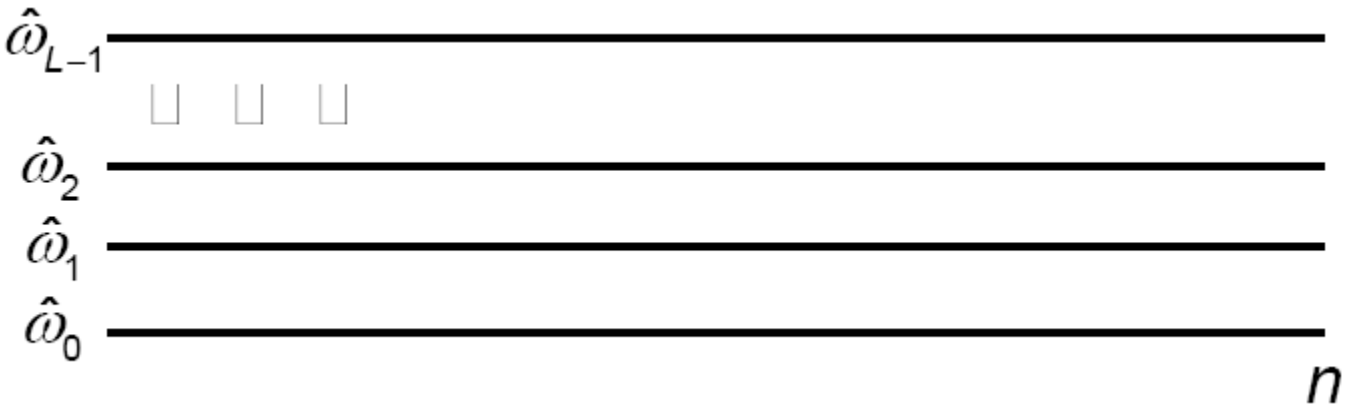
$$\text{DFT: } X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=\hat{n}-L+1}^{\hat{n}} \left(x[m]w[\hat{n}-m] \right) e^{-j\hat{\omega}m}$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \text{DFT}\left(x[m]w[\hat{n}-m] \right)$$

$$0 \le \hat{\omega} < 2\pi,\ \ \hat{n} = 0, R, 2R, \ldots$$

Summary – Modulation/Lowpass Filter

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m}, -\infty < \hat{n} < \infty, 0 \le \hat{\omega} < 2\pi$$



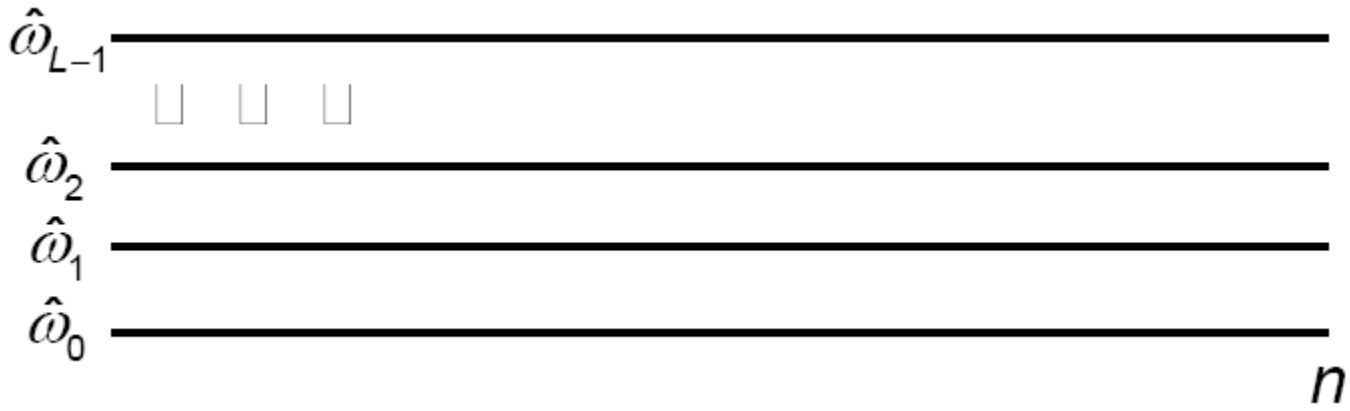
Filter Bank: $X_n(e^{j\hat{\omega}}) = \sum_{m=n-L+1}^{n} \left(x[m]e^{-j\hat{\omega}m}\right)w[n-m]$

$$X_n(e^{j\hat{\omega}}) = \left(x[n]e^{-j\hat{\omega}n}\right)w[n-m]$$

$$= \left(x[n]e^{-j\hat{\omega}n}\right) * w[n]$$

Summary – Bandpass Filter/Demodulation

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m},\ -\infty < \hat{n} < \infty,\ 0 \le \hat{\omega} < 2\pi$$



Filter Bank: $X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} \left(x[n-m]e^{-j\hat{\omega}(n-m)} \right) w[m]$

$$X_n(e^{j\hat{\omega}}) = e^{-j\hat{\omega}n}\left[(w[n]e^{j\hat{\omega}n}) * x[n] \right]$$

STFT Magnitude Only

- for many applications you only need the magnitude of the STFT(not the phase)
- in such cases, the bandpass filter implementation is less complex, since

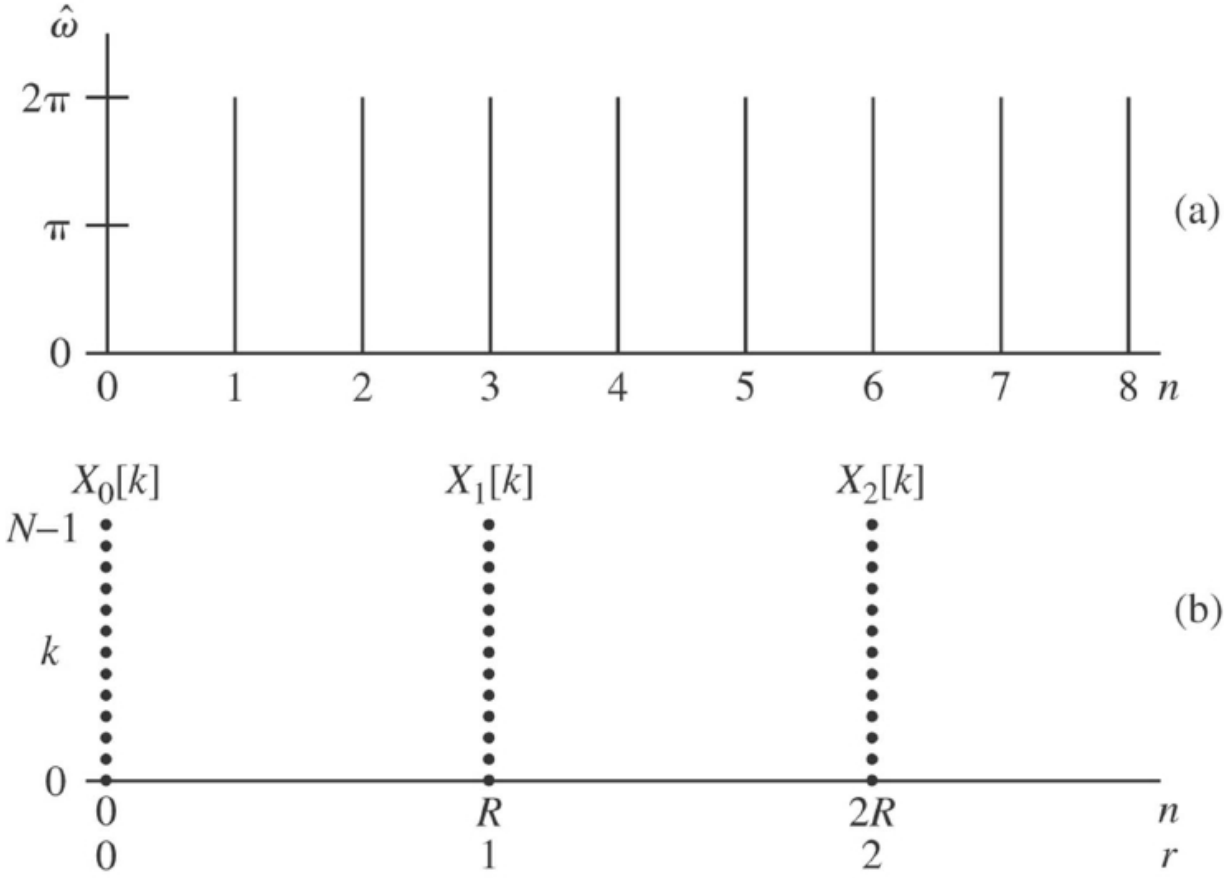
$$|X_n(e^{j\hat{\omega}})| = \left[a_n^2(\hat{\omega}) + b_n^2(\hat{\omega})\right]^{1/2}$$

$$= |\tilde{X}_n(e^{j\hat{\omega}})| = \left[\tilde{a}_n^2(\hat{\omega}) + \tilde{b}_n^2(\hat{\omega})\right]^{1/2}$$

Sampling Rates of STFT

Sampling Rates of STFT

- need to sample STFT in both time and frequency to produce an unaliased representation from which $x(n)$ can be exactly recovered



Sampling Rate in Time

- to determine the sampling rate in time, we take a linear filtering view
  1. $X_n(e^{j\hat{\omega}})$ is the output of a filter with impulse response $\tilde{w}(n)$
  2. $\tilde{W}(e^{j\hat{\omega}})$ is a lowpass response with effective bandwidth of *B* Hertz
- thus the effective bandwidth of $X_n(e^{j\hat{\omega}})$ is *B* Hertz =>$X_n(e^{j\hat{\omega}})$ has to be sampled at a rate of 2*B* samples/second to avoid aliasing

Example: Hamming Window

$$w(n) = 0.54 - 0.46\cos(2\pi n/(L-1)) \quad 0 \le n \le L-1$$
$$= 0 \qquad \text{otherwise}$$

$\Rightarrow B \approx \frac{2F_s}{L}$(Hz); for $L = 400$, $F_s = 10{,}000$ Hz => $B = 50$ Hz => need rate of 100/sec (every 100 samples) for sampling rate in time

Sampling Rate in Frequency

- since $X_n(e^{j\hat{\omega}})$ is periodic in $\hat{\omega}$ with period 2π, it is only necessary to sample over an interval of length 2 π
- need to determine an appropriate finite set of frequencies, $\hat{\omega}_k = 2\pi k / N, k = 0,1,...,N-1$ at which $X_n(e^{j\hat{\omega}})$ must be specified to exactly recover $x(n)$
- use the Fourier transform interpretation of $X_n(e^{j\hat{\omega}})$
  1. if the window $w(n)$ is time-limited, then the inverse transform of $X_n(e^{j\hat{\omega}})$ is time-limited
  2. since the inverse Fourier transform of $X_n(e^{j\hat{\omega}})$ is the signal $x(m)w(n\text{-}m)$ and this signal is of duration *L* samples (the duration of $w(n)$), then according to the sampling theorem $X_n(e^{j\hat{\omega}})$ must be sampled (in frequency) at the set of frequencies $\hat{\omega}_k = 2\pi k / N, k = 0,1,...,N-1, N \geq L$ in order to exactly recover $x(n)$ from $X_n(e^{j\hat{\omega}})$
- thus for a Hamming window of duration *L*=400 samples, we require that the STFT be evaluated at least 400 uniformly spaced frequencies around the unit circle

"Total" Sampling Rate of STFT

- the "total" sampling rate for the STFT is the product of the sampling rates in time and frequency, i.e.,

  $SR = SR(\text{time}) \; x \; SR(\text{frequency})$

  $= 2B \; x \; L$ samples/sec

  B = frequency bandwidth of window (Hz)

  L = time width of window (samples)

- for most windows of interest, B is a multiple of F_S/L, i.e.,

  $B = C \, F_S/L$ (Hz), C=1 for Rectangular Window

  C=2 for Hamming Window

  $SR = 2C \, F_S$ samples/second

- can define an 'oversampling rate' of

  $SR/ F_S = 2C =$ oversampling rate of STFT as compared to conventional sampling representation of $x(n)$

  for RW, $2C=2$; *for HW* $2C=4$ => range of oversampling is 2-4

  this oversampling gives a very flexible representation of the speech signal

Sampling the STFT

- DFT Notation

$$X_r[k] = X_{rR}(e^{j\frac{2\pi}{N}k}) = e^{-j\frac{2\pi}{N}krR}\tilde{X}_r[k]$$

- let $w[-m] \neq 0$ for $0 \leq m \leq L\text{-}1$ (finite duration window with no zero-valued samples)

$$\tilde{X}_r[k] = \sum_{m=0}^{L-1} x[rR+m]w[-m]\, e^{-j\frac{2\pi}{N}km}$$

$$(r \text{ fixed}, \ 0 \leq k \leq N-1)$$

- if $L \leq N$ then (DFT defined with no aliasing => can recover sequence exactly using inverse DFT)

$$x[rR+m]w[-m] = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}_r[k]\, e^{j\frac{2\pi}{N}km}$$

$$(r \text{ fixed}, \ 0 \leq m \leq N-1)$$

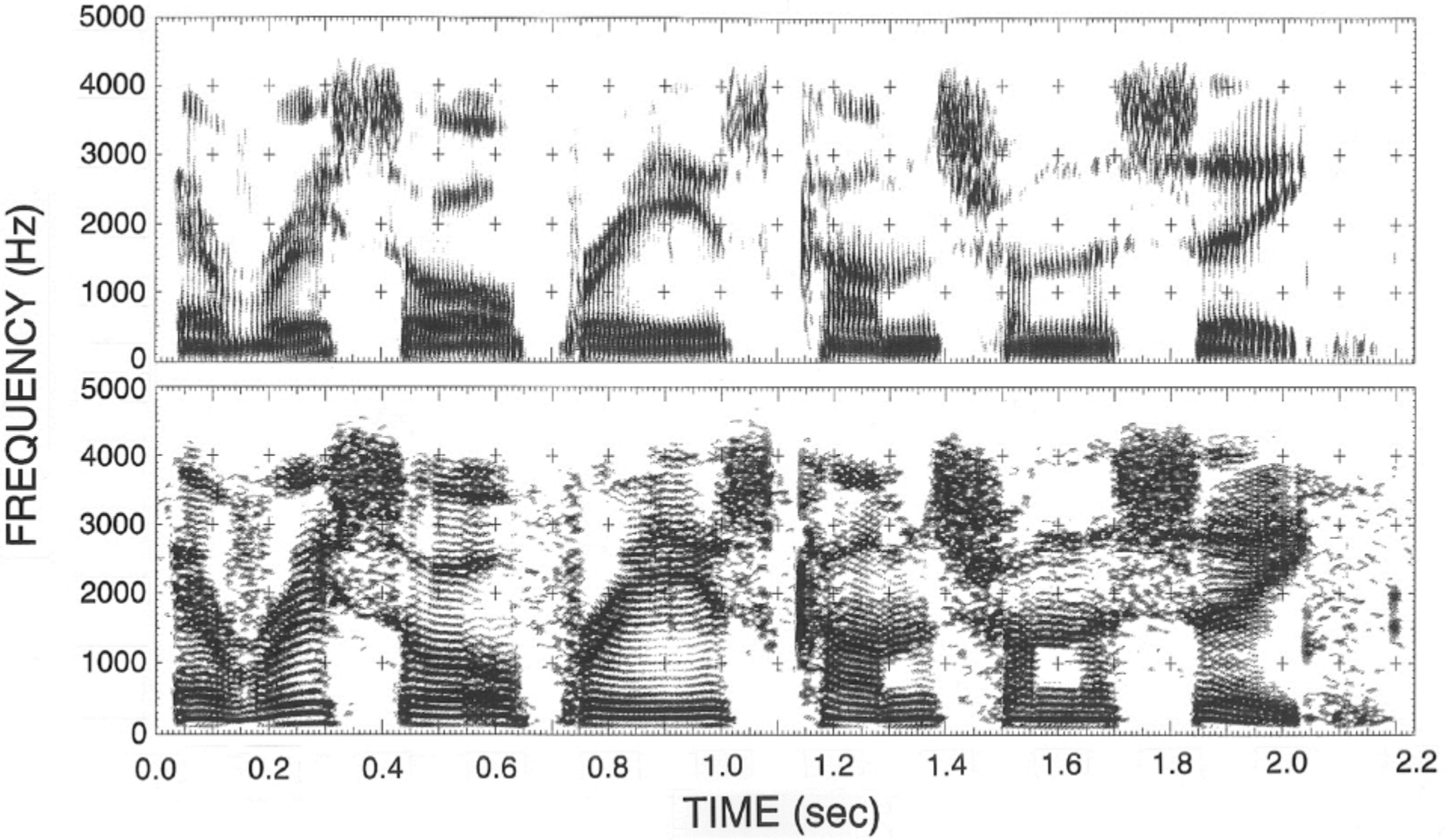
- if $R \leq L$, then all samples can be recovered from $X_r[k]$ ($R > L$ => gaps in sequence)

Spectrographic Displays

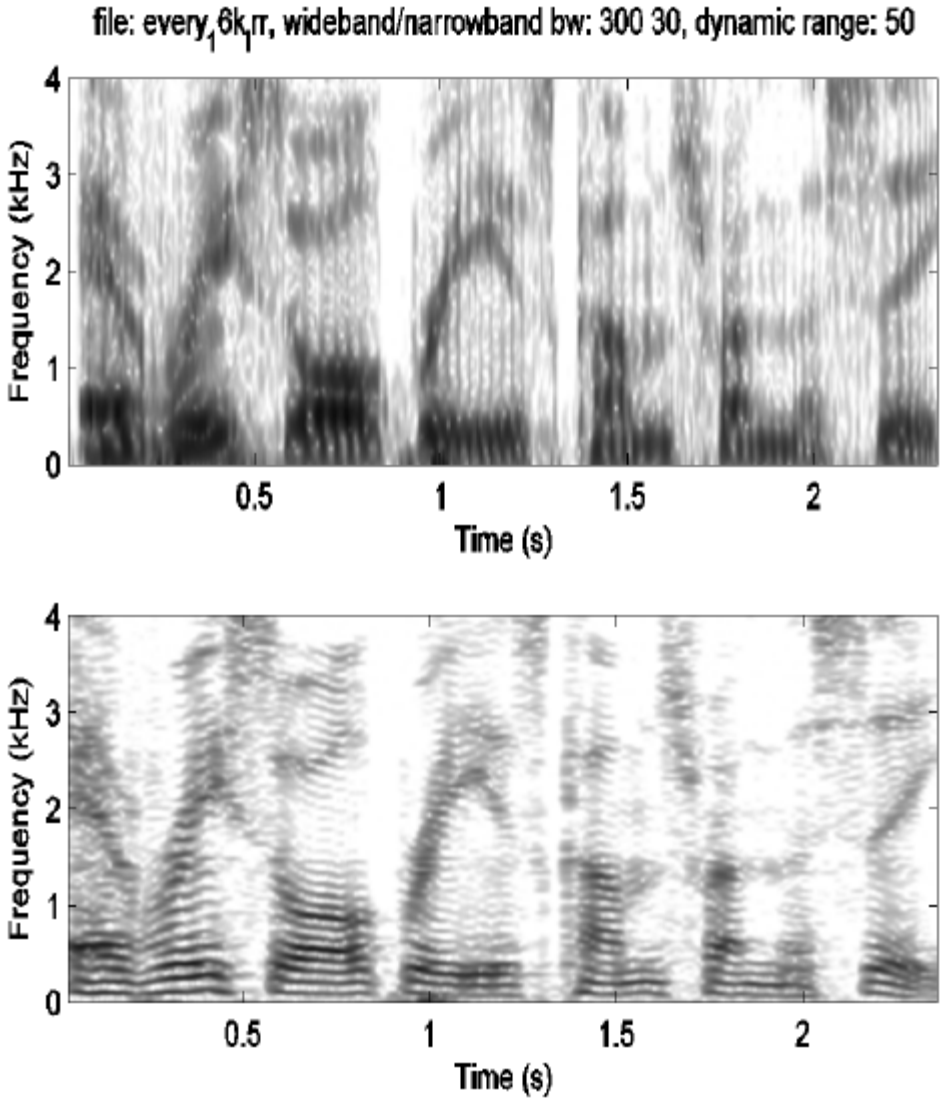
Spectrographic Displays

- Sound Spectrograph-one of the earliest embodiments of the time-dependent spectrum analysis techniques
  - Time-varying average energy in the output of a variable frequency bandpass filter is measured and used as a crude measure of the STFT
  - thus energy is recorded by an ingenious electro-mechanical system on special electrostatic(静电) paper called teledeltos paper(电记录纸)
  - result is a two-dimensional representation of the time-dependent spectrum: with vertical intensity being spectrum level at a given frequency, and horizontal intensity being spectral level at a given time; with spectrum magnitude being represented by the darkness of the marking
  - wide bandpass filters (300 Hz bandwidth) provide good temporal resolution and poor frequency resolution (resolve pitch pulses in time but not in frequency)—called wideband spectrogram
  - narrow bandpass filters (45 Hz bandwidth) provide good frequency resolution and poor time resolution (resolve pitch pulses in frequency, but not in time)—called narrowband spectrogram

Conventional Spectrogram (Every salt breeze comes from the sea)



Digital Speech Spectrograms



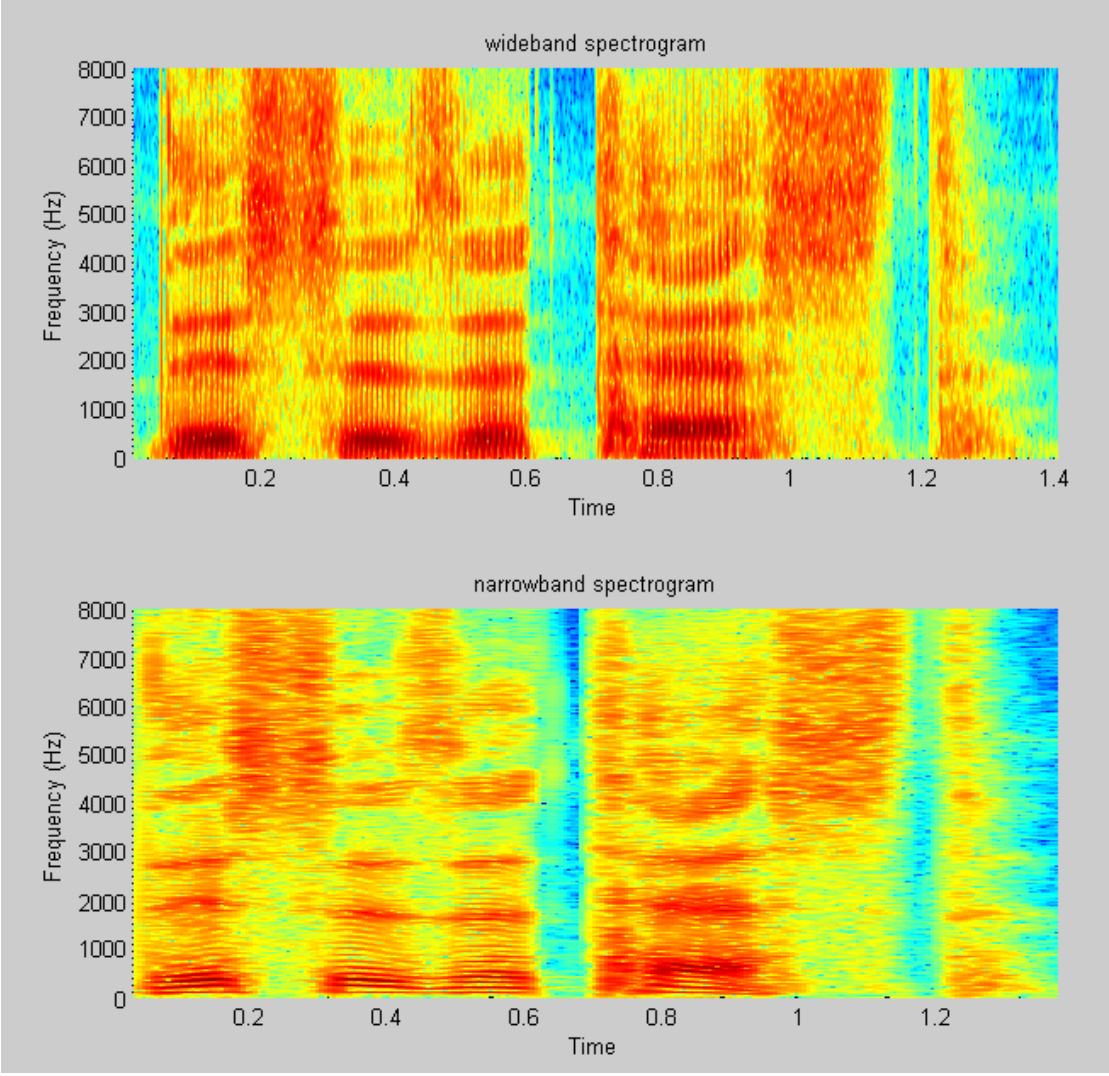
- **wideband spectrogram**
  - follows broad spectral peaks (formants) over time
  - resolves most individual pitch periods as vertical striations since the IR of the analyzing filter is comparable in duration to a pitch period
  - what happens for low pitch males—high pitch females
  - for unvoiced speech there are no vertical pitch striations
- **narrowband spectrogram**
  - individual harmonics are resolved in voiced regions
  - formant frequencies are still in evidence
  - usually can see fundamental frequency
  - unvoiced regions show no strong structure

Digital Speech Spectrograms

- Speech Parameters ("This is a test"): 
  - sampling rate: 16 kHz
  - speech duration: 1.406 seconds
  - speaker: male
- Wideband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 6 msec (96 samples)
  - analysis window shift: 0.625 msec (10 samples)
  - FFT size: 512
- Narrowband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 60 msec (960 samples)
  - analysis window shift: 6 msec (96 samples)
  - FFT size: 1024

- Matlab Example

Digital Speech Spectrograms

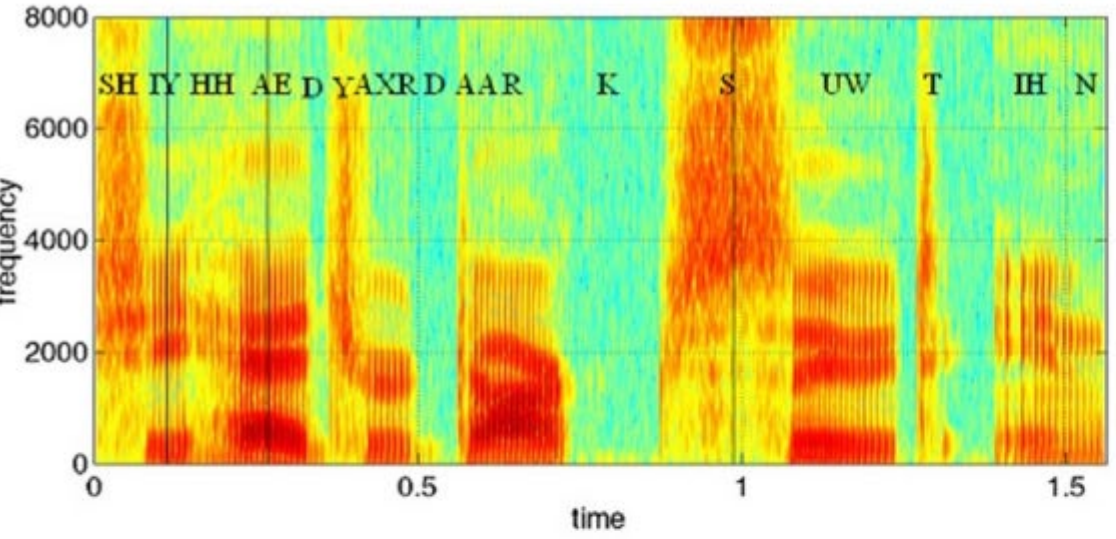


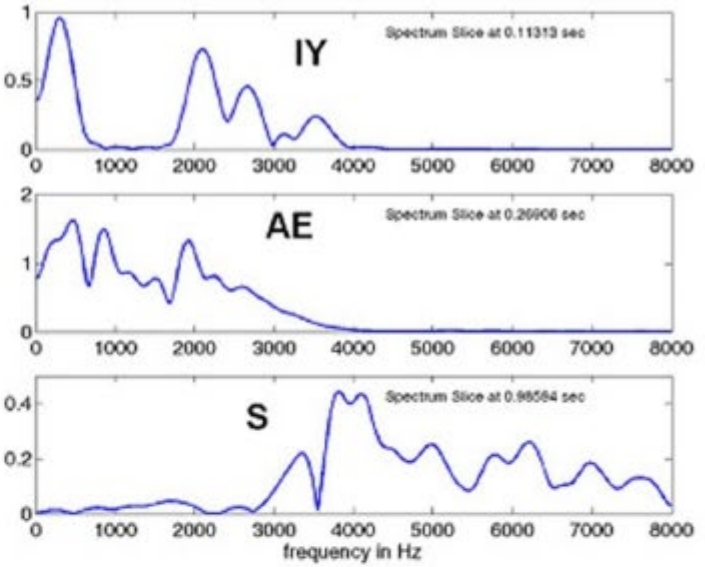
6 msec (96 samples) window

60 msec (960 sample) window

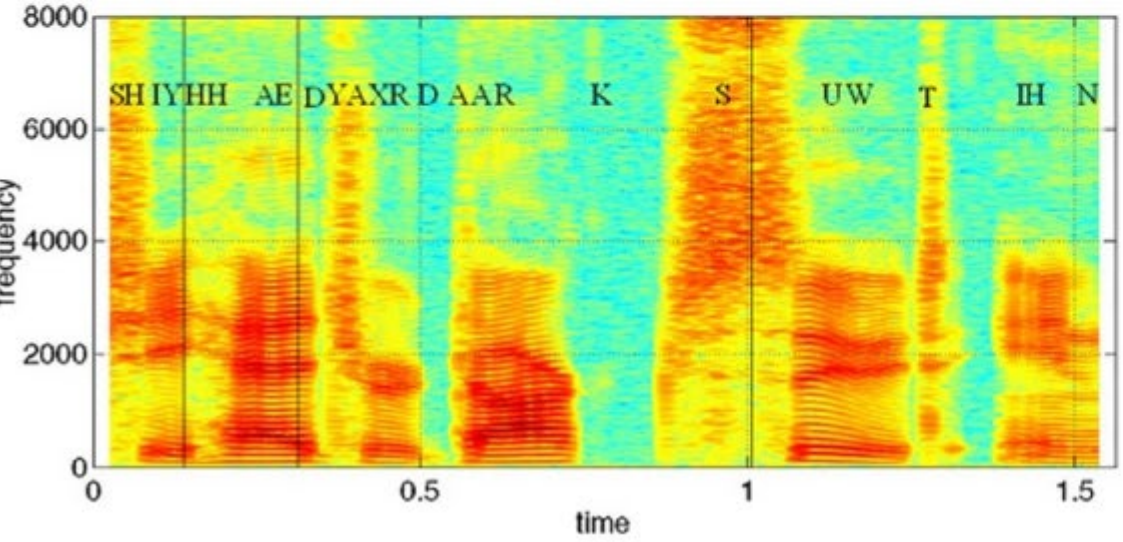
Spectrogram - Male

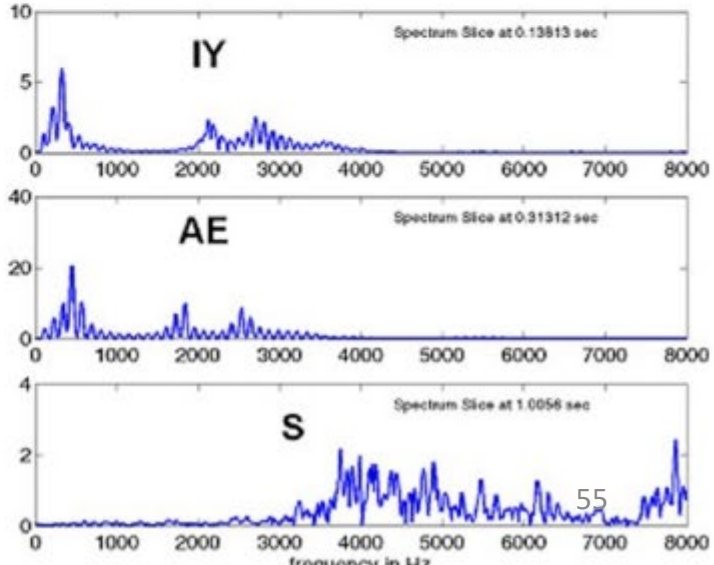
nfft=1024, *L*=80, *R*=5 "She had your dark suit in."





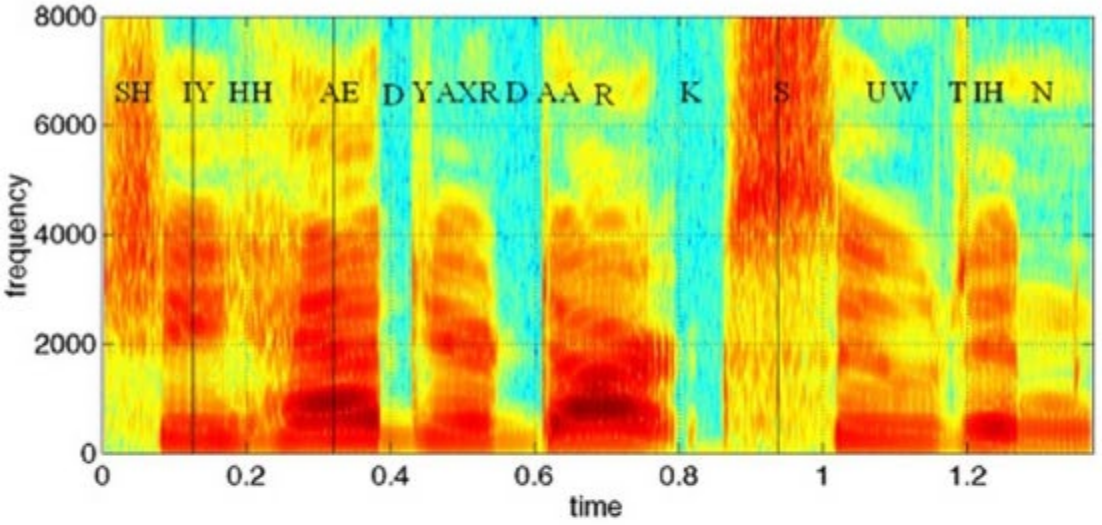
nfft=1024, *L*=800,*R* = 10



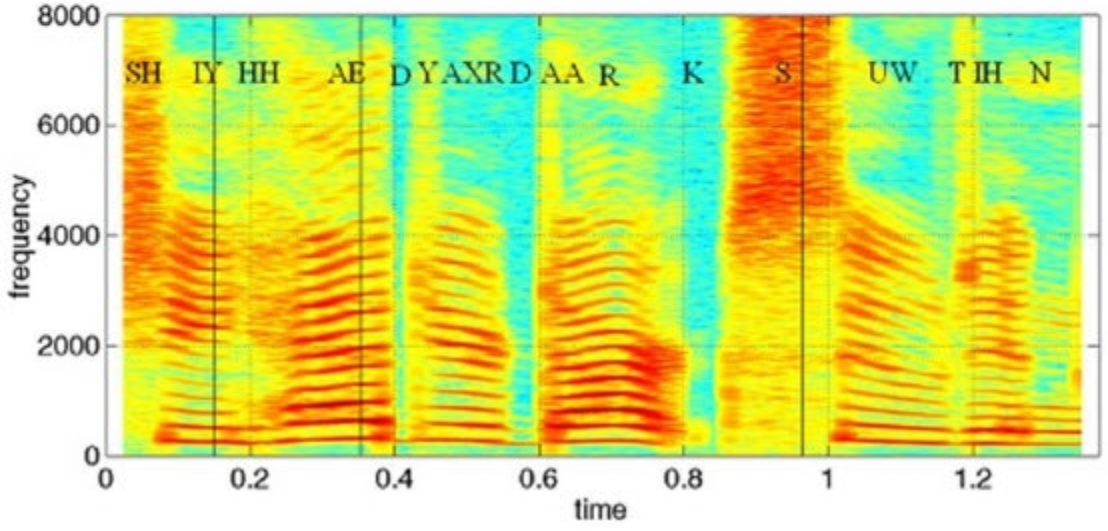


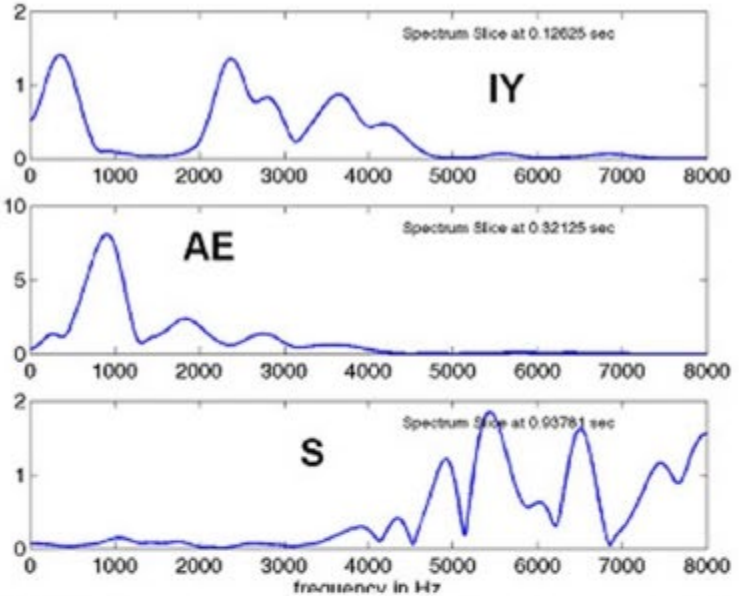
Spectrogram - Female

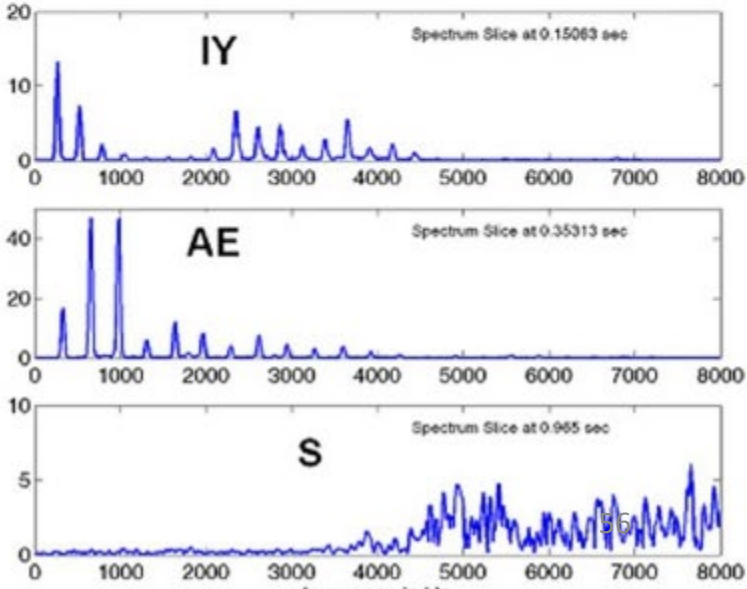
nfft=1024, *L*=80, *R*=5 “She had your dark suit in.”



nfft=1024, *L*=800,*R* = 10







A Summary on Introduced STFS Methods

Method #1

- since $X_{\hat{n}}(e^{j\hat{\omega}})$ is the normal Fourier transform of the window sequence $w(\hat{n}-m)x(m)$, then

$$w(\hat{n}-m)\,x(m)=\frac{1}{2\pi}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}$$

- with the requirement that $w(0)\neq 0$, the sequence $x(\hat{n})$ can be recovered from $X_{\hat{n}}(e^{j\hat{\omega}})$, if $X_{\hat{n}}(e^{j\hat{\omega}})$ is known for every value of $\hat{n}$ and for all $\hat{\omega}$

$$x(\hat{n})=\frac{1}{2\pi w(0)}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

Method #2

- $X_{\hat{n}}(e^{j\hat{\omega}})$ can be recovered from its sample version

$$X_r[k] = X_{rR}(e^{j\frac{2\pi}{N}k}) = e^{-j\frac{2\pi}{N}krR}\tilde{X}_r[k]$$

if $R \le F_s/2B$ and $N \ge L$, where *B* is the window bandwidth

Example: Hamming Window

$$w(n) = 0.54 - 0.46\cos(2\pi n/(L-1)) \quad 0 \le n \le L-1$$
$$= 0 \quad \text{otherwise}$$

$\Rightarrow B \approx \frac{2F_s}{L}$(Hz); for $L = 400$, $F_s = 10{,}000$ Hz => $B = 50$ Hz => need rate of 100/sec (every 100 samples) for sampling rate in time

Method #3

- DFT Notation

$$X_r[k] = X_{rR}(e^{j\frac{2\pi}{N}k}) = e^{-j\frac{2\pi}{N}krR}\tilde{X}_r[k]$$

- let $w[-m] \neq 0$ for $0 \leq m \leq L\text{-}1$ (finite duration window with no zero-valued samples)

$$\tilde{X}_r[k] = \sum_{m=0}^{L-1} x[rR+m]w[-m]\, e^{-j\frac{2\pi}{N}km}$$

$$(r \text{ fixed}, \ 0 \leq k \leq N-1)$$

- if $L \leq N$ then (DFT defined with no aliasing => can recover sequence exactly using inverse DFT)

$$x[rR+m]w[-m] = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}_r[k]\, e^{j\frac{2\pi}{N}km}$$

$$(r \text{ fixed}, \ 0 \leq m \leq N-1)$$

- if $R \leq L$, then all samples can be recovered from $X_r[k]$ ($R > L$ => gaps in sequence)

Overlap Addition (OLA) Method

Overlap Addition (OLA) Method

- based on normal FT interpretation of short-time spectrum

$$X_{\hat{n}}(e^{j\omega_k}) \xleftarrow{DFT/IDFT} y_{\hat{n}}(m) = x(m)w(\hat{n}-m)$$

- can reconstruct *x*(*m*) by computing IDFT of $X_{\hat{n}}(e^{j\omega_k})$ and dividing out the window (assumed non-zero for all samples)
- this process gives *L* signal values of *x*(*m*) for each window => window can be moved by *L* samples and the process repeated
- This procedure is theoretically valid with *R*<=*L*<=*N*
- Not practical since small changes in $X_{rR}(e^{j\omega_k})$ will be amplified by dividing the inverse DFT by the window

Overlap Addition (OLA) Method

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

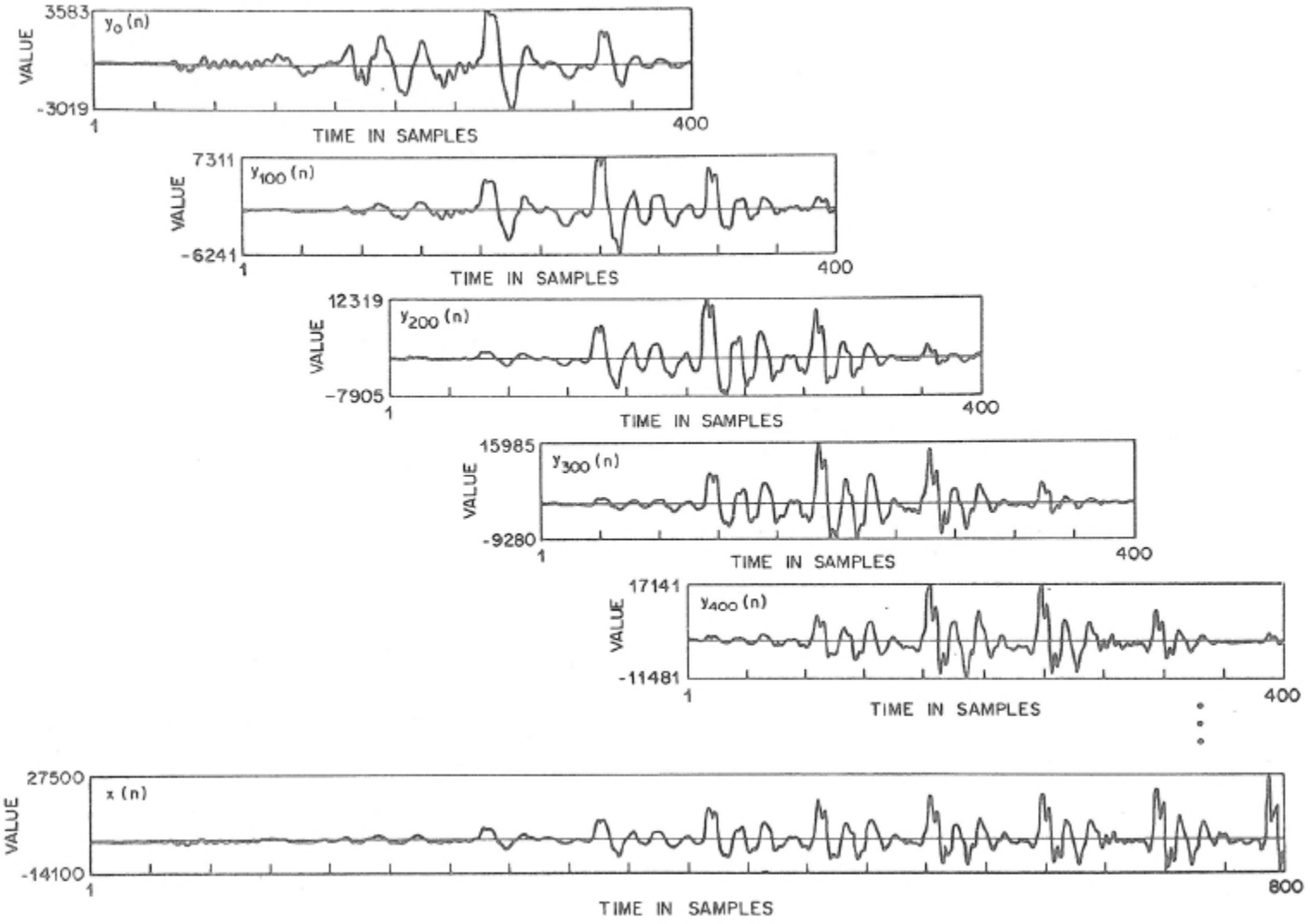
- summation is for overlapping analysis sections
- for each value of *m* where $X_m(e^{j\omega_k})$ is measured, do an inverse FT to give

$$y_m(n) = Lx(n)w(m-n) \text{ (where } L \text{ is the size of the FT)}$$
$$y(n) = \sum_m y_m(n) = Lx(n)\sum_m w(m-n)$$

- The condition for exact reconstruction of *x*[*n*] is

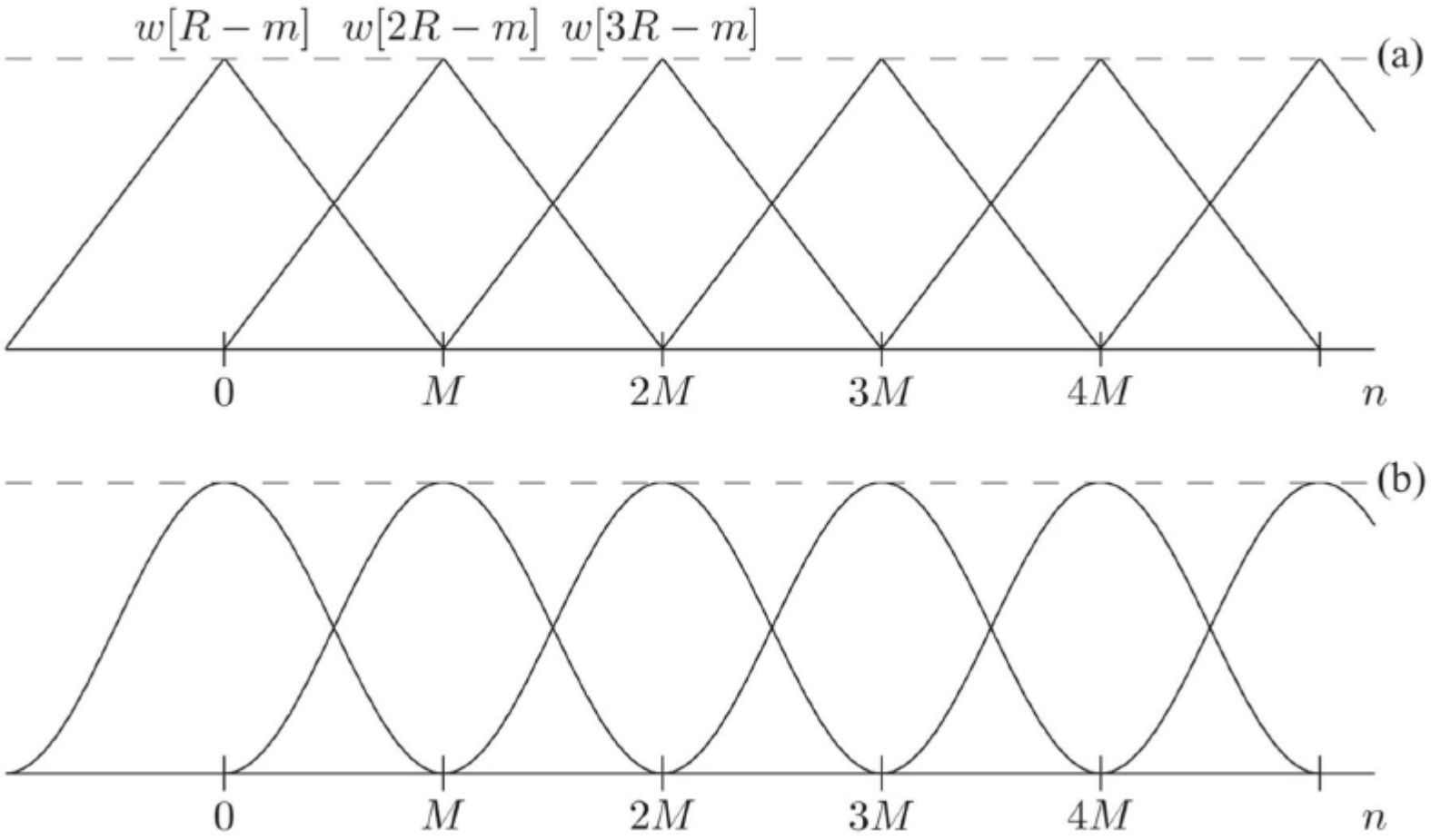
$$\widetilde{w}[n] = \sum_{r=-\infty}^{\infty} w[rR - n] = C$$

Overlap Addition (OLA) Method



Overlap Addition of Bartlett and Hann Windows

$L = 2M+1 \quad R = M$



Spectral Condition

$$w[n] \Leftrightarrow W(e^{j\omega})$$

$$w[-n] \Leftrightarrow W^*(e^{j\omega})$$

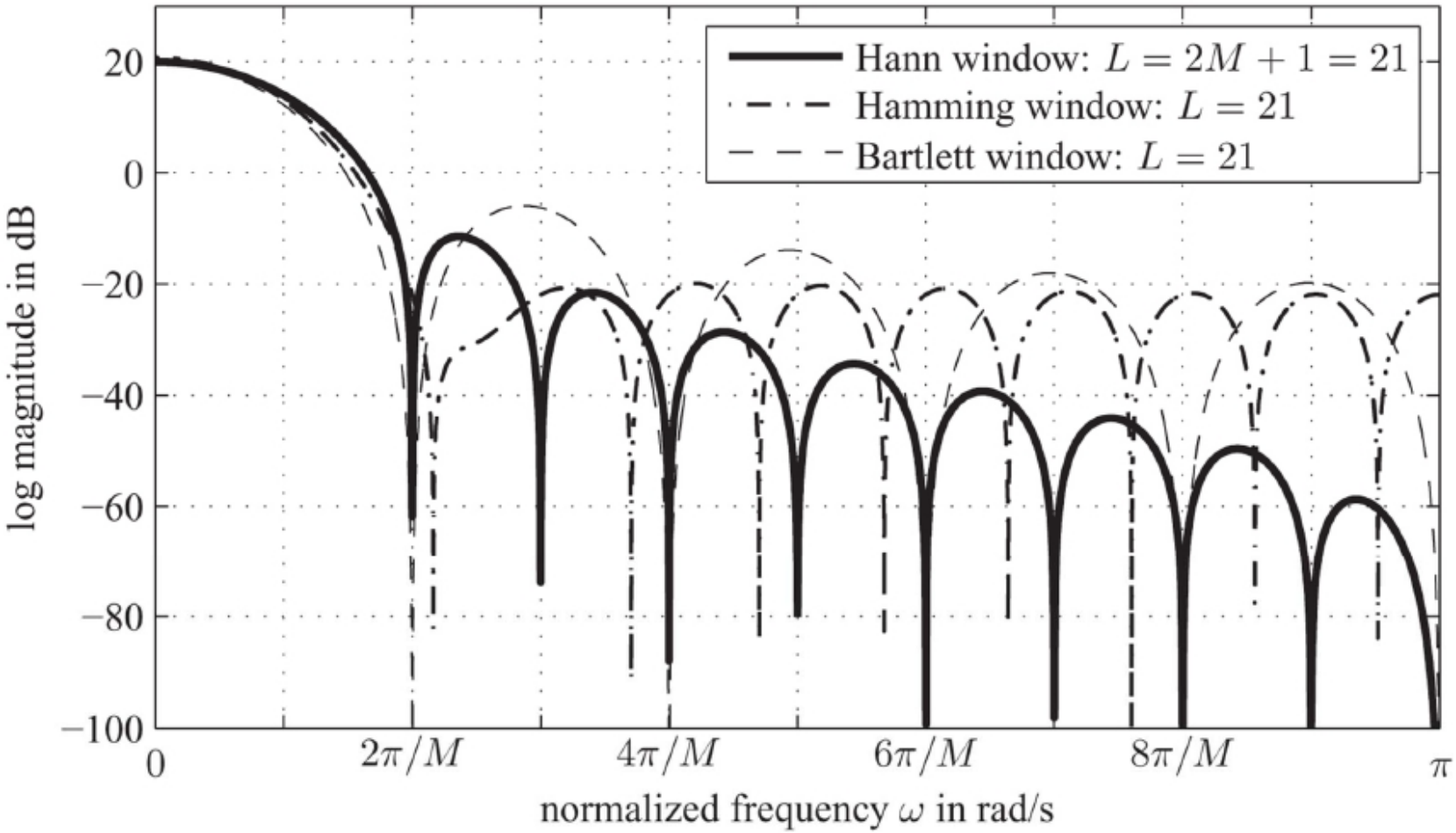
$$\widetilde{w}[n] = \sum_{r=-\infty}^{\infty} w[rR-n] \Leftrightarrow W^*(e^{j(2\pi k/R)})$$

$$\widetilde{w}[n] = \sum_{r=-\infty}^{\infty} w[rR-n] = \frac{1}{R}\sum_{k=0}^{R-1} W^*(e^{j(2\pi k/R)})e^{j(2\pi k/R)n}$$

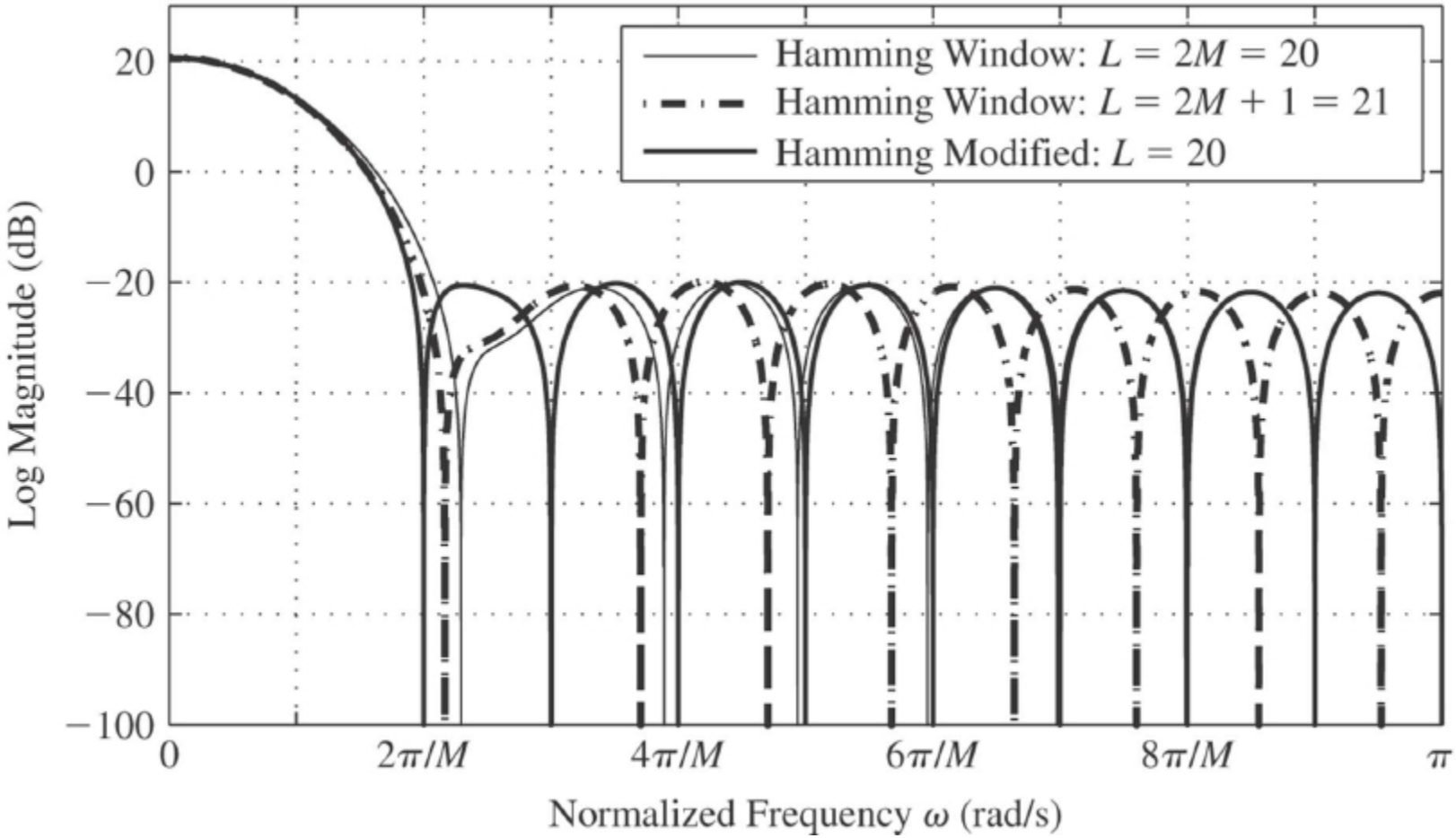
One sufficient condition for perfect reconstruction is:

$$\left|W^*(e^{j(2\pi k/R)})\right| = \left|W(e^{j(2\pi k/R)})\right| = 0, \;\; k = 1, 2, \ldots, R-1$$

Window Spectra

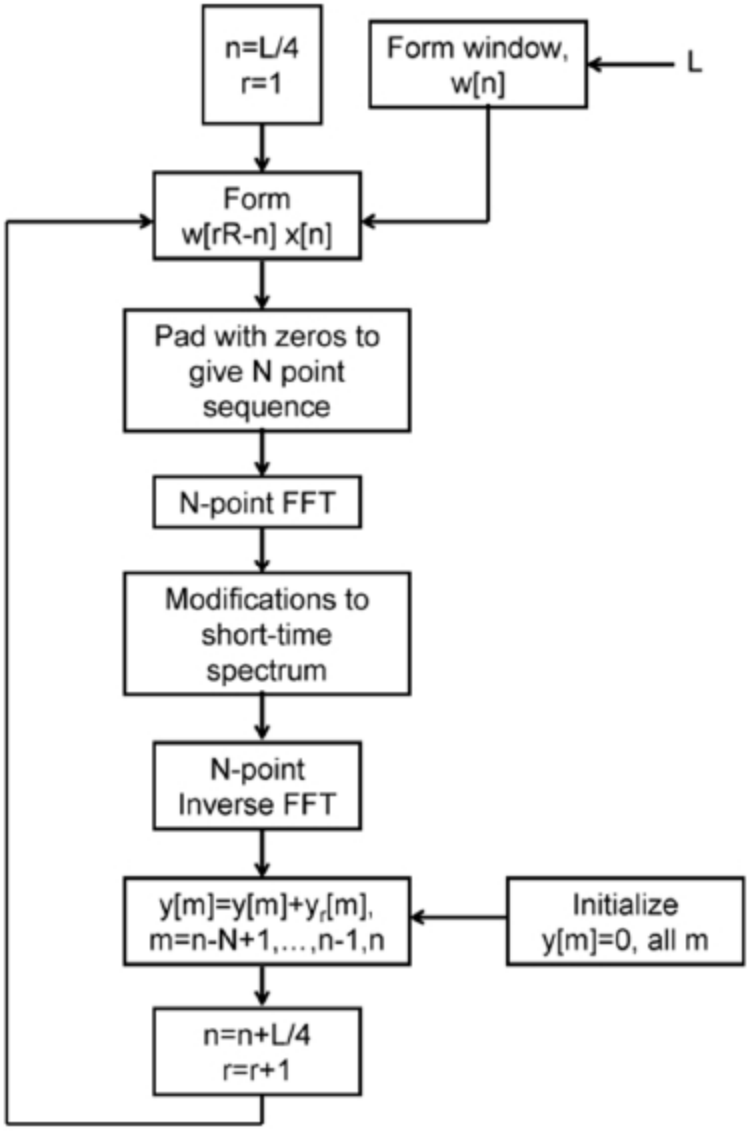


Hamming Window Spectra



- DTFTs of even-length, odd-length and modified odd-to-even length Hamming windows
- Odd-to-even: truncate from L = 2M+1 to L = 2M by simply zeroing the last sample; zeros spaced at 2π/R give perfect reconstruction using OLA

Overlap Addition (OLA) Method



- *w(n)* is an L-point Hamming window with *R=L/4*
- assume *x(n)=0* for *n<0*
- time overlap of 4:1 for HW
- first analysis section begins at *n=L/4*

Overlap Addition (OLA) Method

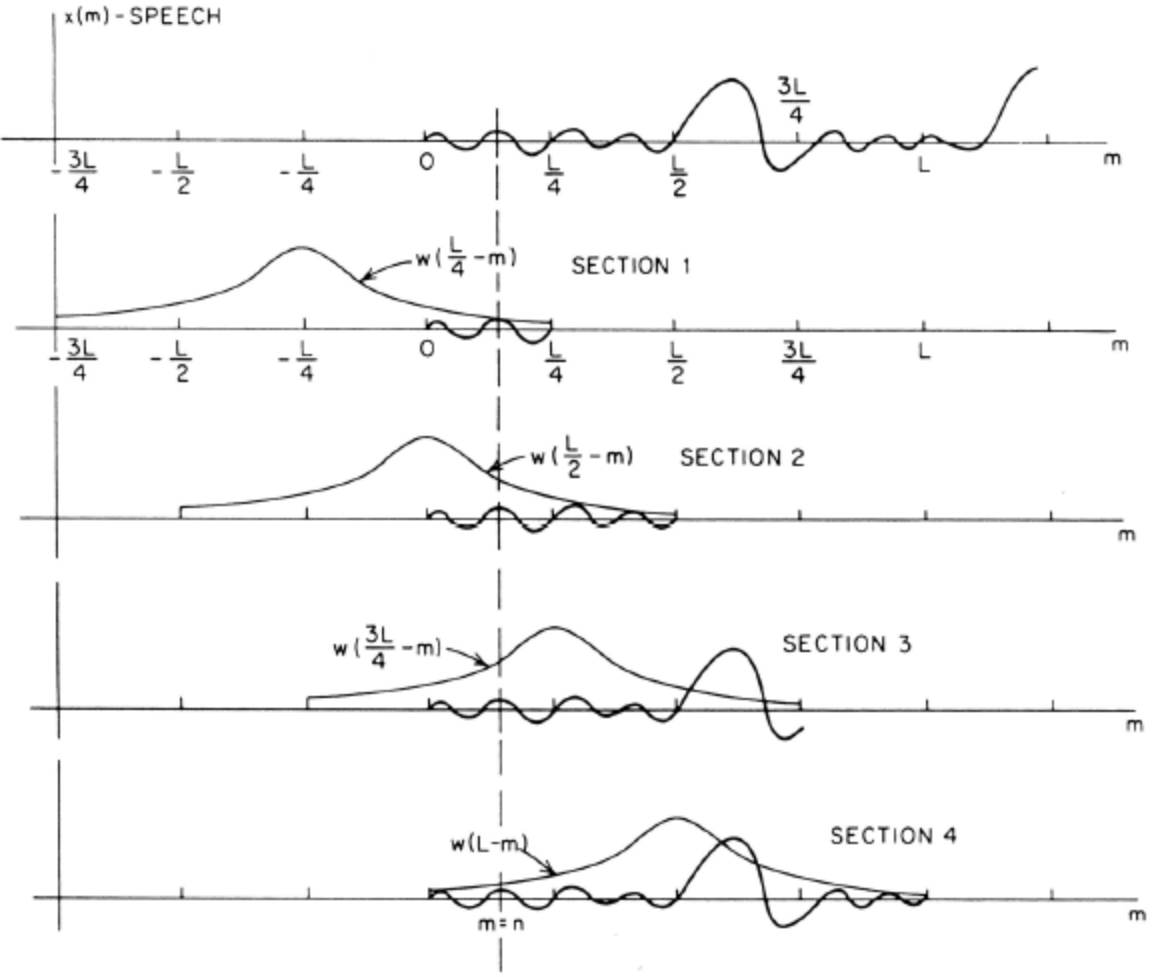


Fig. 6.17 Reconstruction procedure for $w(n)$ using an L-point Hamming window.

- 4-overlapping sections contribute to each interval
- *N*-point FFT's done using *L* speech samples, with *N-L* zeros padded at end to allow modifications without significant aliasing effects
- for a given value of *n*

$$y(n)=x(n)w(R-n)+x(n)w(2R-n)+x(n)w(3R-n)+x(n)w(4R-n)=x(n)[w(R-n)+w(2R-n)+w(3R-n)+w(4R-n)]=x(n)\ W(e^{j0})/R$$

Filter Bank Summation
(FBS)

Filter Bank Summation

- the filter bank interpretation of the STFT shows that for any frequency ω_k , $X_n(e^{j\omega_k})$ is a lowpass representation of the signal in a band centered at ω_k ($n = \hat{n}$ for FBS)

$$X_n(e^{j\omega_k}) = e^{-j\omega_k n} \sum_{m=-\infty}^{\infty} x(n-m) w_k(m)\, e^{j\omega_k m}$$

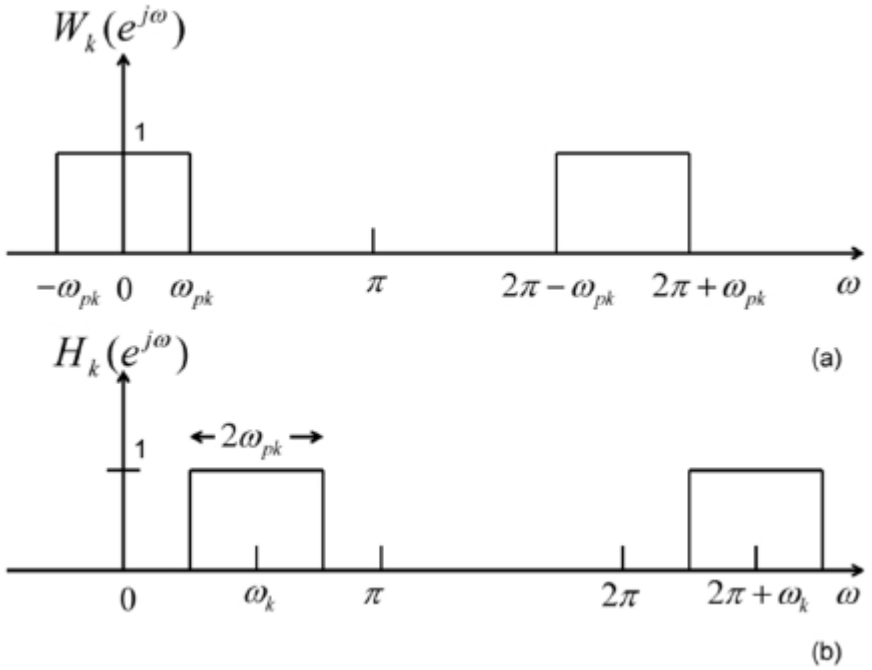
where $w_k(m)$ is the lowpass window used at frequency ω_k

Filter Bank Summation

- define a bandpass filter and substitute it in the equation to give

$$h_k(n) = w_k(n)\, e^{j\omega_k n}$$

$$X_n(e^{j\omega_k}) = e^{-j\omega_k n} \sum_{m=-\infty}^{\infty} x(n-m) h_k(m)$$



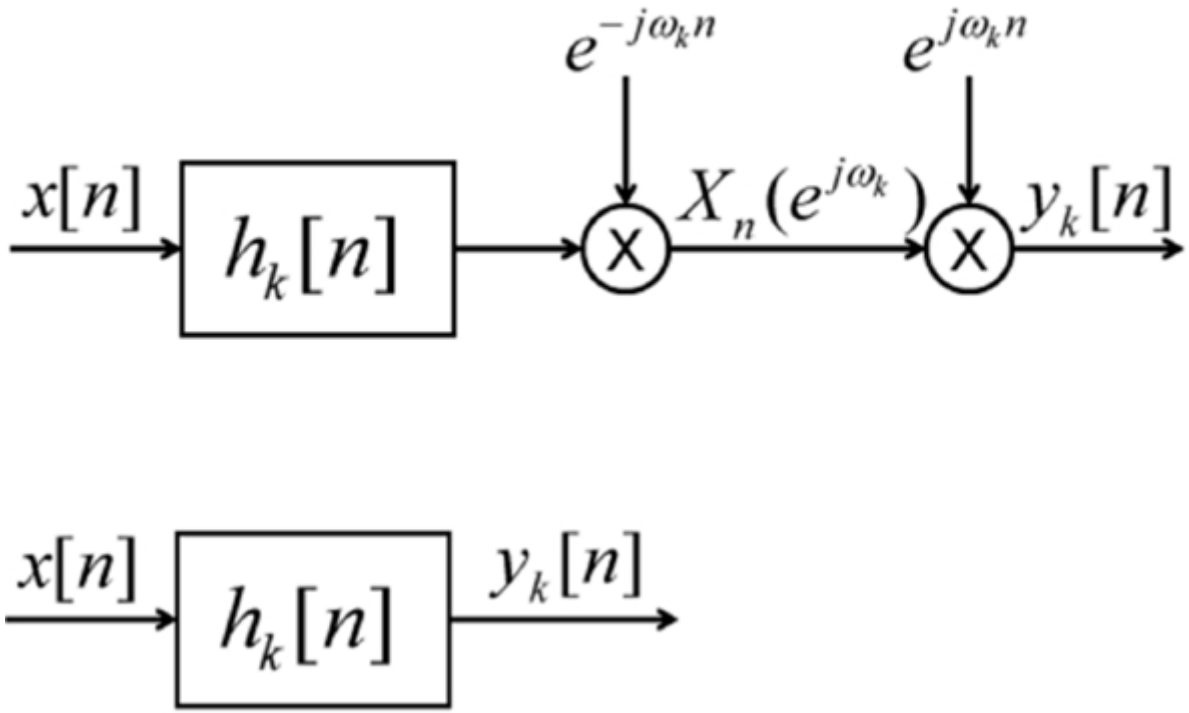
Filter Bank Summation

- thus $X_n(e^{j\omega_k})$ is obtained by bandpass filtering *x(n)* followed by modulation with the complex exponential $e^{-j\omega_k n}$. We can express this in the form

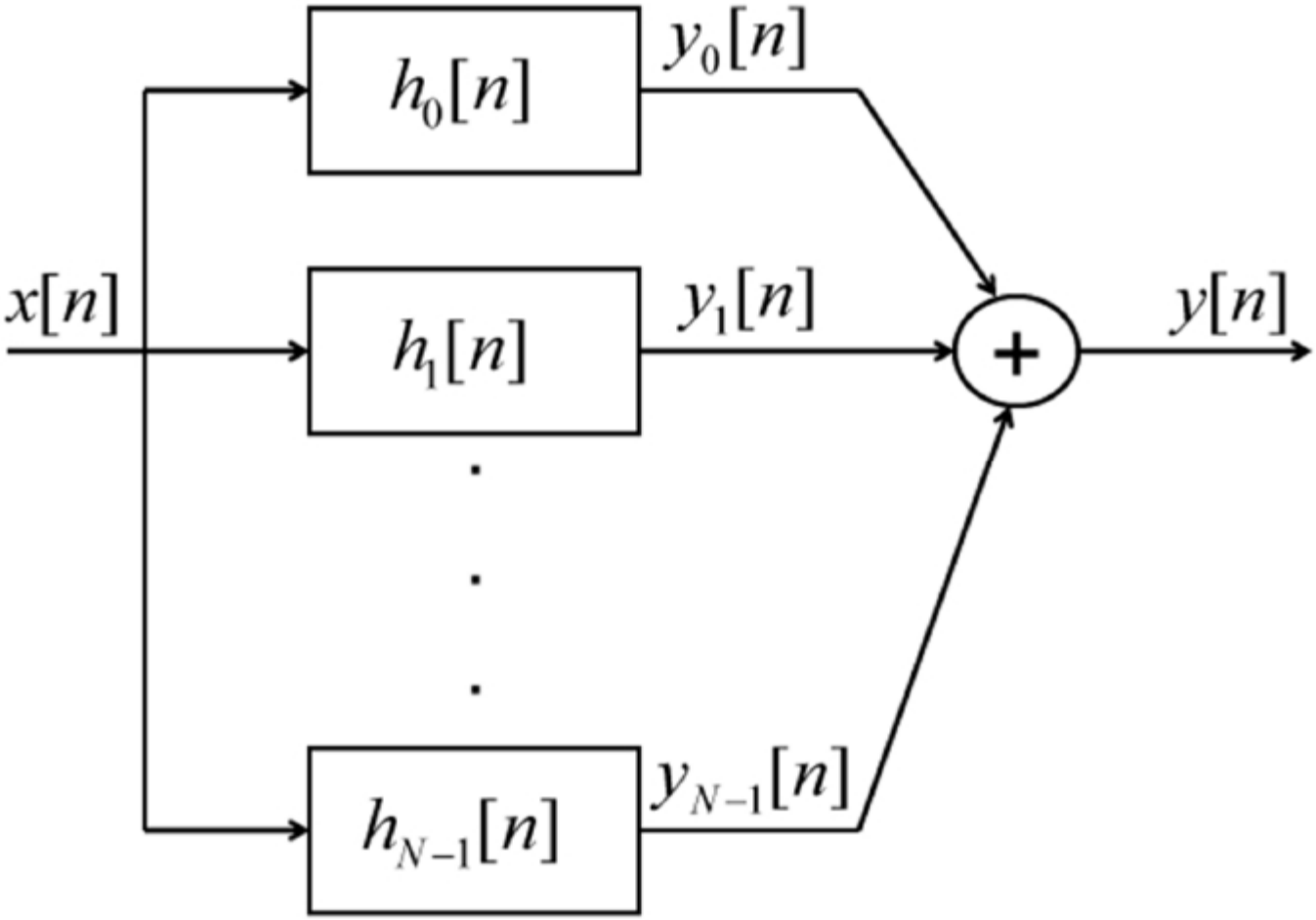
$$y_k(n) = X_n(e^{j\omega_k})e^{j\omega_k n} = \sum_{m=-\infty}^{\infty} x(n-m)h_k(m)$$

- thus $y_k(n)$ is the output of a bandpass filter with impulse response $h_k(n)$

Filter Bank Summation



Filter Bank Summation



Filter Bank Summation

- consider a set of *N* bandpass filters, uniformly spaced, so that the entire frequency band is covered

$$\omega_k = \frac{2\pi k}{N},\ k = 0,1,\ldots,N-1$$

- also assume window the same for all channels, i.e.,

$$w_k(n) = w(n),\quad k = 0,1,\ldots,N-1$$

- if we add together all the bandpass outputs, the composite response is

$$\tilde{H}(e^{j\omega}) = \sum_{k=0}^{N-1} H_k(e^{j\omega}) = \sum_{k=0}^{N-1} W(e^{j(\omega-\omega_k)})$$

- if $W(e^{j\omega_k})$ is properly sampled in frequency ($N \geq L$), where *L* is the window duration, then it can be shown that

$$\frac{1}{N}\sum_{k=0}^{N-1} W(e^{j(\omega-\omega_k)}) = w(0) \quad \forall\,\omega$$

FBS Formula

Proof of FBS Formula

- derivation of FBS formula

$$w(n) \xleftarrow{FT\ /\ IFT} W(e^{j\omega})$$

- if $W(e^{j\omega})$ is sampled in frequency at uniformly spaced points, the inverse discrete Fourier transform of the sampled version of $W(e^{j\omega_k})$ is (recall that sampling $\Rightarrow$ multiplication $\Leftrightarrow$ convolution $\Rightarrow$ aliasing)

$$\frac{1}{N}\sum_{k=0}^{N-1} W(e^{j\omega_k})e^{j\omega_k n} = \sum_{r=-\infty}^{\infty} w(n+rN)$$

- an aliased version of *w(n)* is obtained.

Proof of FBS Formula

- If *w(n)* is of duration *L* samples, then

$$w(n) = 0, \; n < 0, \;\; n \geq L$$

- and no aliasing occurs due to sampling in frequency of $W(e^{j\omega})$
- In this case if we evaluate the aliased formula for n = 0, we get

$$\frac{1}{N}\sum_{k=0}^{N-1} W(e^{j\omega_k}) = w(0)$$

- the FBS formula is seen to be equivalent to the formula above, since (according to the sampling theorem) any set of N uniformly spaced samples of $W(e^{j\omega})$ is adequate.

Filter Bank Summation

- the impulse response of the composite filter bank system is

$$\tilde{h}(n) = \sum_{k=0}^{N-1} h_k(n) = \sum_{k=0}^{N-1} w(n) e^{j\omega_k n} = N w(0) \delta(n)$$

- thus the composite output is

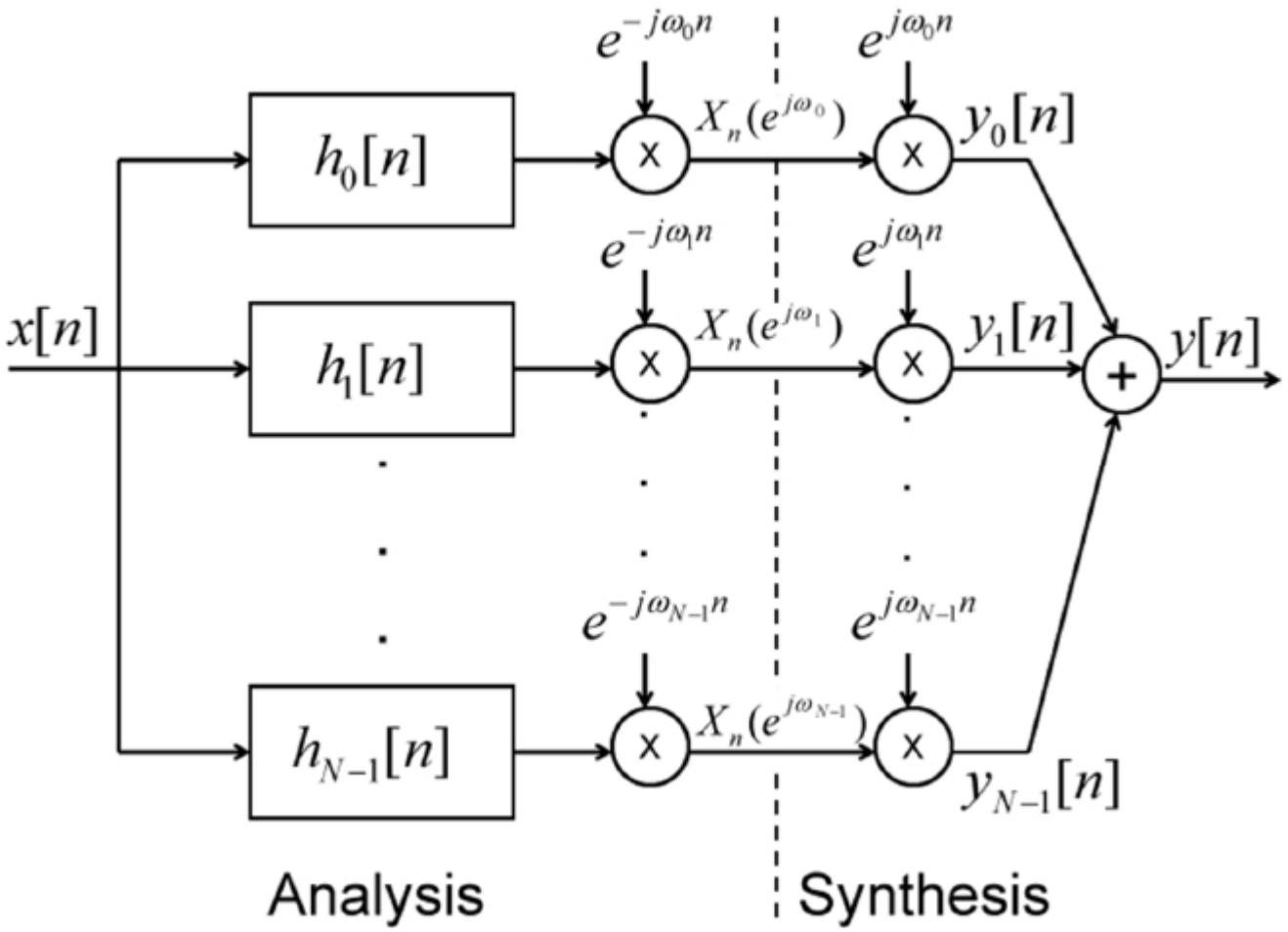
$$y(n) = x(n) * \tilde{h}(n) = N w(0) x(n)$$

- thus for FBS method, the reconstructed signal is

$$y(n) = \sum_{k=0}^{N-1} y_k(n) = \sum_{k=0}^{N-1} X_n(e^{j\omega_k}) e^{j\omega_k n} = N w(0) x(n)$$

if $X_n(e^{j\omega_k})$ is sampled properly in frequency, and is independent of the shape of *w(n)*

Filter Bank Summation



FBS Reconstruction

- the composite impulse response for the FBS system is

$$\tilde{h}(n) = \sum_{k=0}^{N-1} w(n) e^{j\omega_k n} = w(n) \sum_{k=0}^{N-1} e^{j\omega_k n}$$

- defining a composite of the terms being summed as

$$p(n) = \sum_{k=0}^{N-1} e^{j\omega_k n} = \sum_{k=0}^{N-1} e^{j2\pi kn/N}$$

- we get for $\tilde{h}(n)$

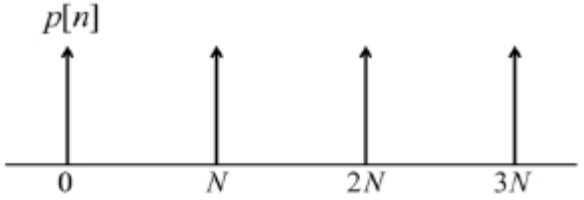
$$\tilde{h}(n) = w(n)\, p(n)$$

- it is easy to show that *p(n)* is a periodic train of impulses of the form

$$p(n) = N \sum_{r=-\infty}^{\infty} \delta(n - rN)$$

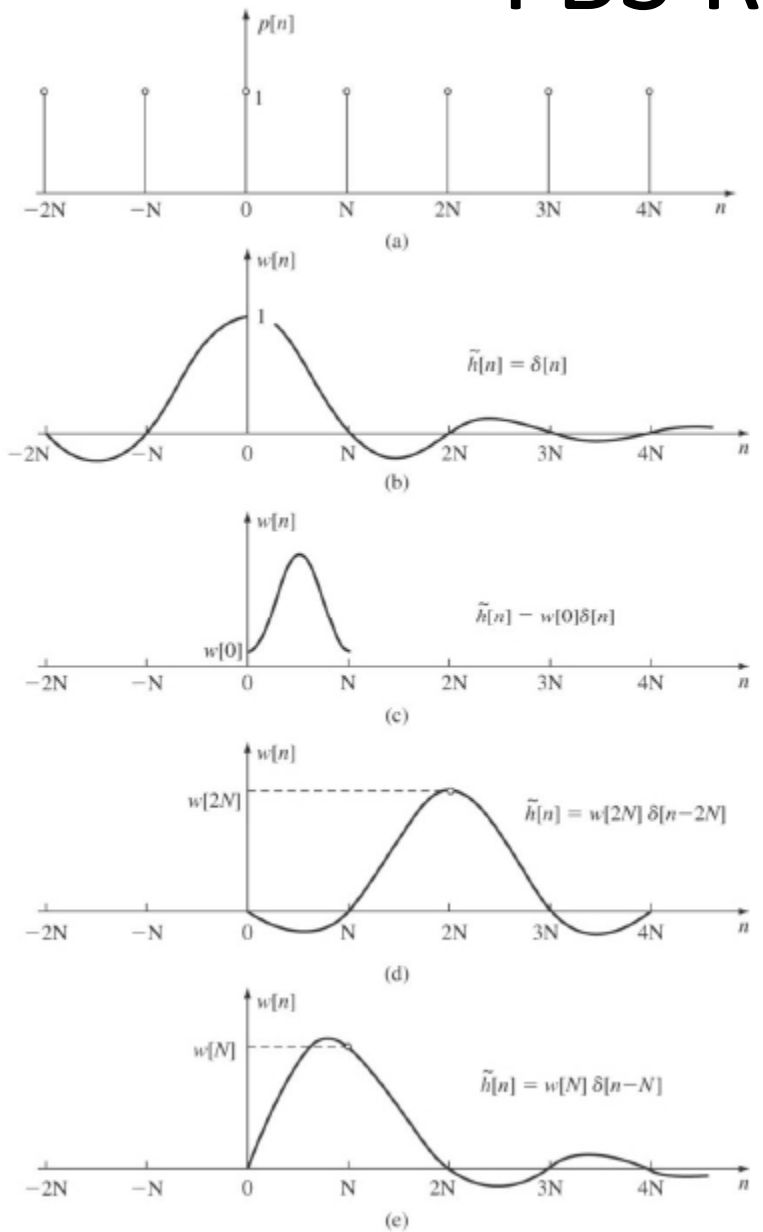
- giving for $\tilde{h}(n)$ the expression

$$\tilde{h}(n) = N \sum_{r=-\infty}^{\infty} w(rN)\, \delta(n - rN)$$



- thus the composite impulse response is the window sequence sampled at intervals of *N* samples

FBS Reconstruction



impulse response of ideal lowpass filter with cutoff frequency π/N

- for ideal LPF we have

$$w(n) = \frac{\sin(\pi n / N)}{\pi n},\ w(rN) = \frac{\sin(\pi r)}{\pi rN} = \frac{1}{N}\delta(r)$$

giving $\tilde{h}(n) = \delta(n)$

- other cases where perfect reconstruction is obtained

1. $w(n)$ is of finite length $L \le N$ and causal (no images)
2. $w(n)$ has length $> N$ and has the property

$$w(n) = 1/N, \quad \text{for } n = r_0 N$$
$$= 0 \qquad \text{for } n = rN\ (r \ne r_0,\ r = 0, \pm 1, \pm 2, \ldots)$$

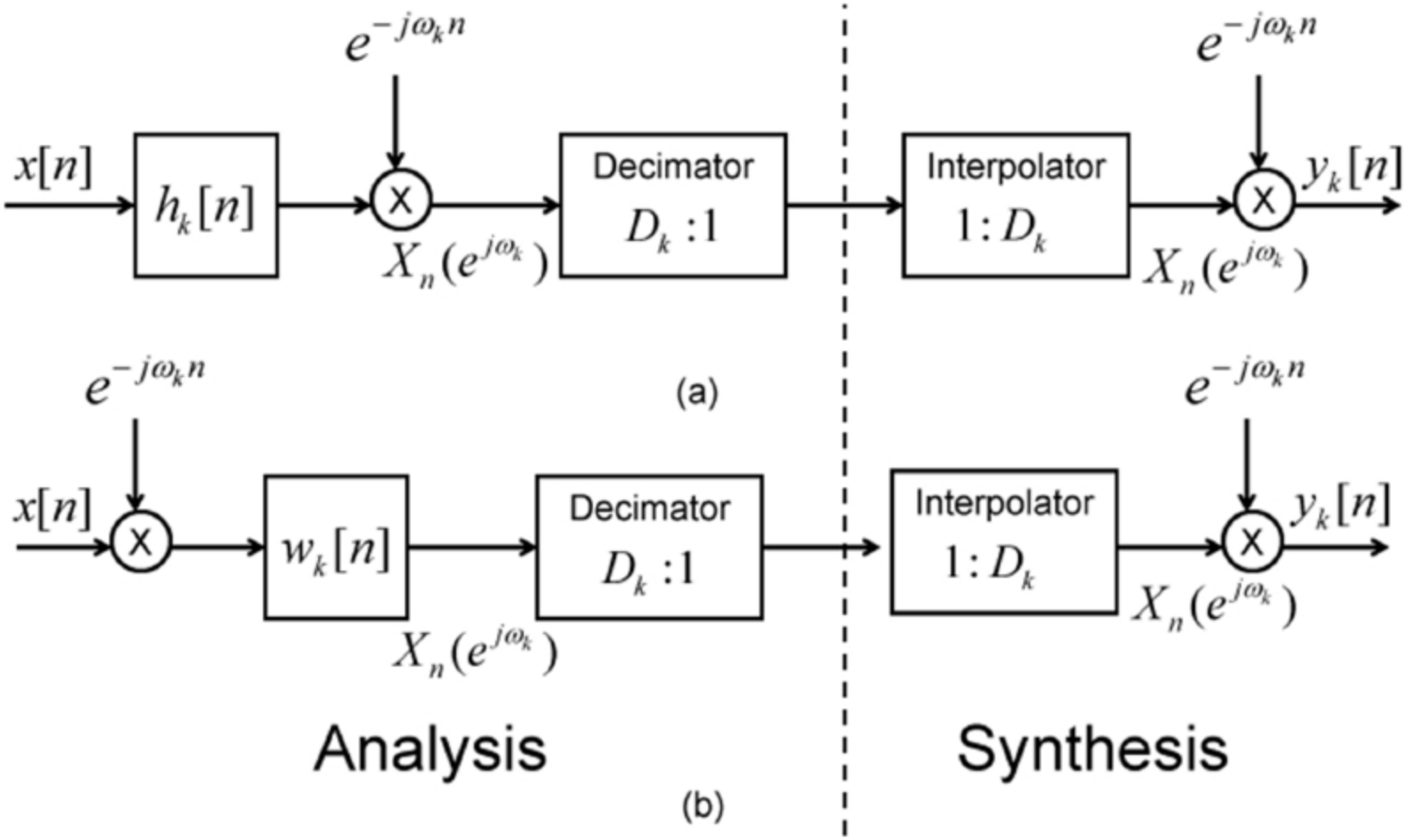
giving $\tilde{h}(n) = p(n)w(n) = \delta(n - r_0 N)$

$$\tilde{H}(e^{j\omega}) = e^{-j\omega r_0 N} \Rightarrow y(n) = x(n - r_0 N)$$

Summary of FBS Reconstruction

- for perfect reconstruction using FBS methods
  1. *w(n)* does not need to be either time-limited or frequency-limited to exactly reconstruct *x(n)* from $X_n(e^{j\omega_k})$
  2. *w(n)* just needs equally spaced zeros, spaced *N* samples apart for theoretically perfect reconstruction
- exact reconstruction of the input is possible with a number of frequency channels less than that required by the sampling theorem
- key issue is how to design digital filters that match these criteria

Practical Implementation of FBS



FBS and OLA Comparisons

FBS and OLA Comparisons

- filter bank summation method $\xleftrightarrow{\text{duals}}$ overlap addition method
  - one depends on sampling relation in frequency
  - one depends on sampling relation in time
- FBS requires sampling in frequency be such that the window transform $W(e^{j\omega})$ obeys the relation

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

- OLA requires that sampling in time be such that the window obeys the relation

$$\sum_{r=-\infty}^{\infty} w(rR-n) = W(e^{j0})/R \quad \text{any } n$$

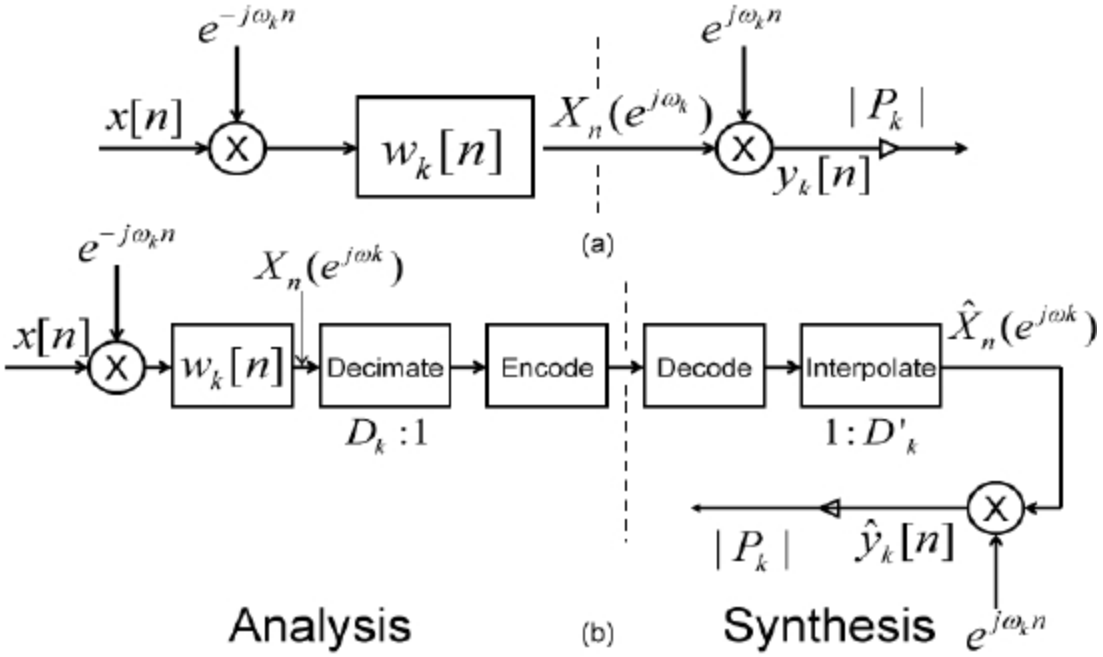
- the key to Short-Time Fourier Analysis is the ability to modify the short-time spectrum via quantization, noise enhancement, signal enhancement, speed-up/slow-down, etc) and recover an "unaliased" modified signal

Applications of STFT

Applications of STFT

- vocoders => voice coders, code speech at rates much lower than waveform coders
- removal of additive noise
- de-reverberation
- speed-up and slow-down of speech for speed learning, aids for the handicapped

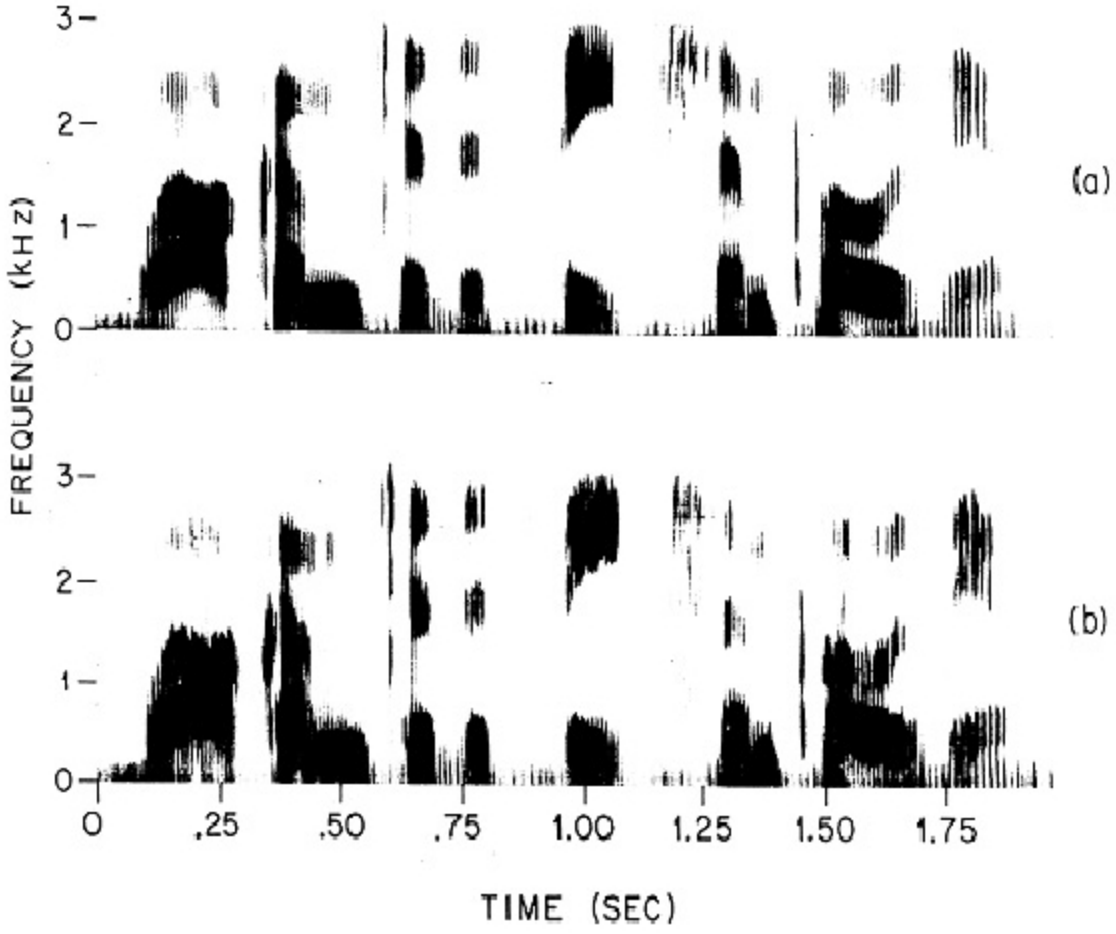
Coding of STFT



- elements of STFT
  1. set of $\{\omega_k\}$ chosen to cover frequency range of interest
  2. $w_k(n)$-set of lowpass analysis windows
  3. P_k-set of complex gains to make composite frequency response as close to ideal as possible

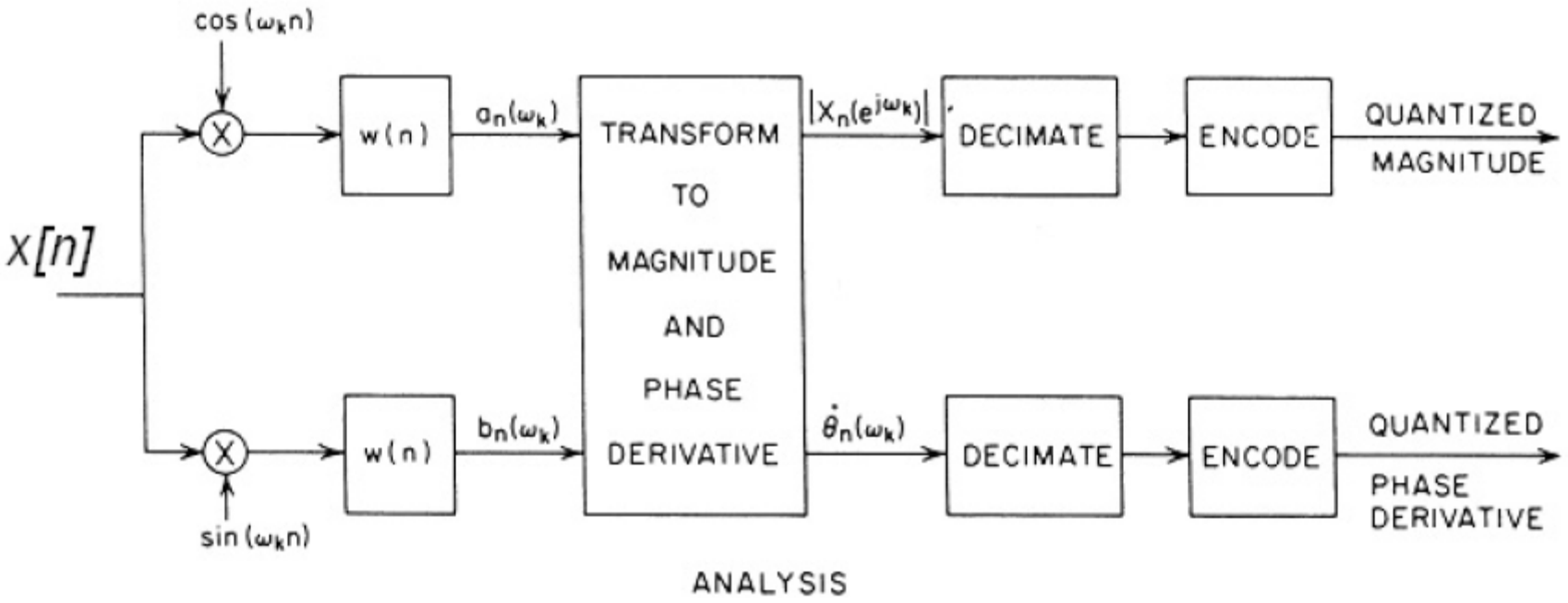
=> goal is to sample STFT at rates lower than $x(n)$

Coding of STFT



- non-uniform coding and quantization
- 28 channels
- 100/sec SR (gives small amount of aliasing)
- coding log magnitude and phase using 3 bits for log magnitude and 4 bits for phase for channels 1-10; and 2 bits for log magnitude and 3 bits for phase for channels 11-28
- total rate of 16 Kbps

The Phase Vocoder



- used for speed-up and slow-down of speech
- speed-up: divide center frequency and phase derivative by *q*
- slow-down: multiply center frequency and phase derivative by *q*

Examples of Rate Changes in Speech

- Female Speaker
  - Original rate
  - Speeded up
  - Speeded up more
  - Slowed down
  - Slowed down more

Modify sampling rate
+30%

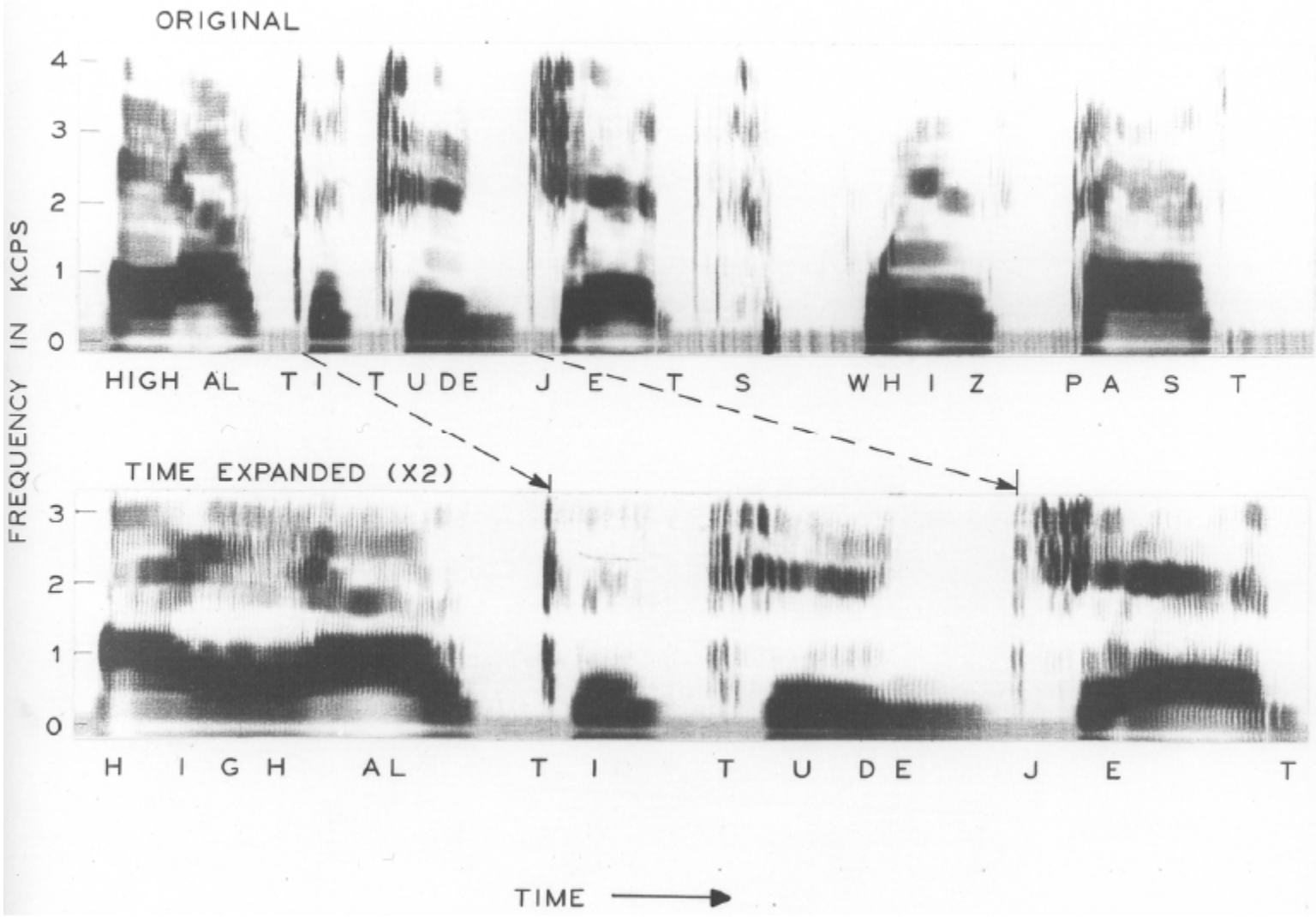
-30%

- Male Speaker
  - Original rate
  - Speeded up
  - Speeded up more
  - Slowed down
  - Slowed down more

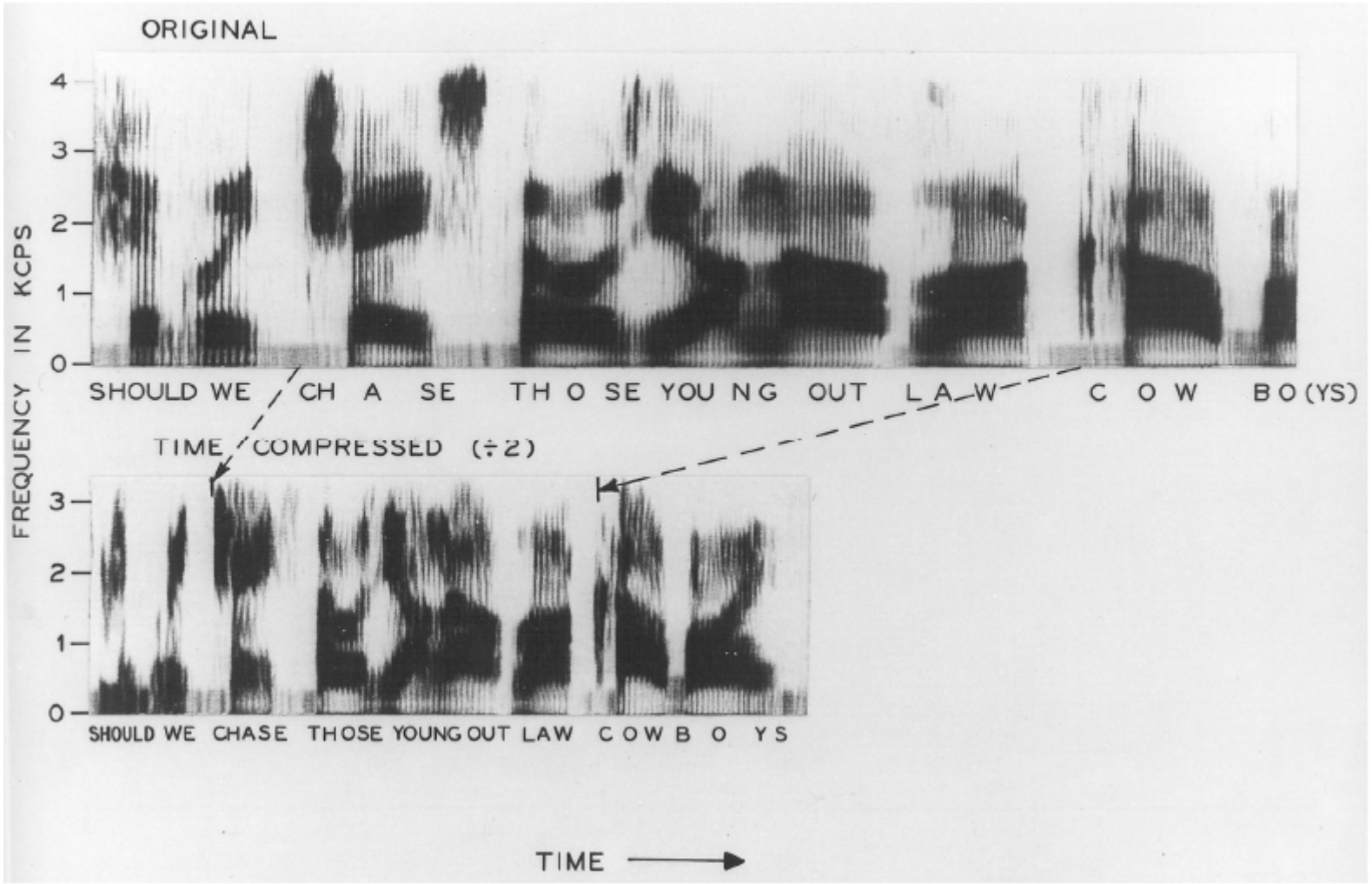
Modify sampling rate
+30%

-30%

Phase Vocoder Time Expanded



Phase Vocoder Time Compressed

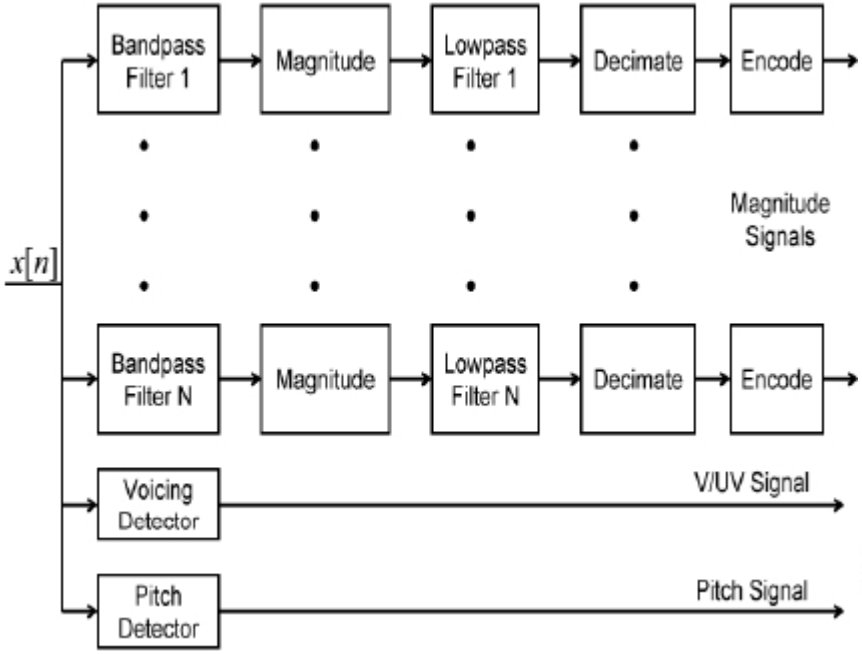


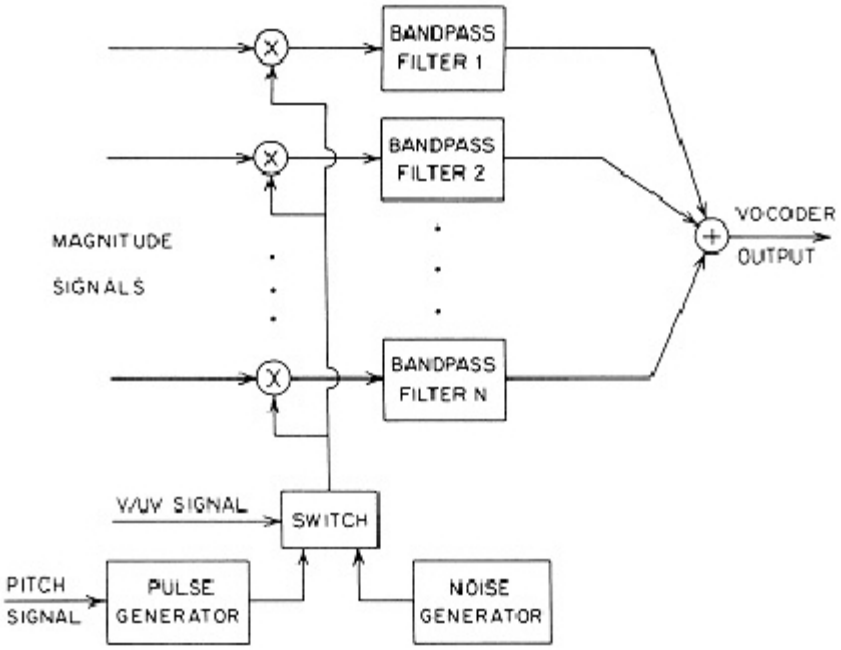
Channel Vocoder

- interpret STFT so that each channel can be thought of as a bandpass filter with center frequency ω_k
- magnitude of STFT can be approximated by envelope detection on the BPF output
- **analyzer**-bank of channels; need excitation info (the phase component) => V/UV detector, pitch detector
- **synthesizer**-channel signal control channel amplitude; excitation signals control detailed structure of output for a given channel; V/UV choice of excitation source

=> **highly reverberant speech because of total lack of control of composite filter bank response**

Channel Vocoder





- 1200-9600 bps
- 600 bps for pitch and V/UV
- easy to modify pitch, timing

Channel Vocoder